MODELO PREDICTIVO PARA LA IDENTIFICACIÓN DE ACTIVIDADES DE LA VIDA DIARIA (ADL) EN AMBIENTES INDOOR USANDO TÉCNICAS DE CLASIFICACIÓN BASADAS EN MACHINE LEARNING

JOHANNA KARINA GARCIA RESTREPO

CORPORACIÓN UNIVERSIDAD DE LA COSTA – CUC

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN Y ELECTRÓNICA

MAESTRIA EN GESTION DE LAS TECNOLOGÍAS DE INFORMACIÓN Y LA COMUNICACIÓN

BARRANQUILLA

2020

MODELO PREDICTIVO PARA LA IDENTIFICACIÓN DE ACTIVIDADES DE LA VIDA DIARIA (ADL) EN AMBIENTES INDOOR USANDO TÉCNICAS DE CLASIFICACIÓN BASADAS EN MACHINE LEARNING

JOHANNA KARINA GARCIA RESTREPO

Trabajo de grado para optar el Titulo de Magister en Gestión de las Tecnologías de la Información y la Comunicación

Tutores

Ing. Emiro De-La-Hoz-Franco, PhD

Ing. Paola Ariza-Colpas, MSc

CORPORACIÓN UNIVERSIDAD DE LA COSTA – CUC

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN Y ELECTRÓNICA

MAESTRIA EN GESTION DE LAS TECNOLOGÍAS DE INFORMACIÓN Y LA COMUNICACIÓN

BARRANQUILLA

2020

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

Firma del Presidente del Jurado

_____

Jurado

_____

Jurado

Agradecimietos

Quiero agradecer a Dios por permitirme hacer realizar una meta más en mi vida.

Agradecer a la Universidad de la Costa por brindarme la oportunidad de hacer parte de esta maravillosa institución.

A mi tutor Emiro de la Hoz, el cuale me orientó e hizo posible este proyecto.

A mis compañeros de la Maestría, por su apoyo, confianza, por compartir sus conocimientos y fortalecer los lazos de amistad.

A mi familia, por ser un pilar fundamental en todo este proceso y apoyarme en cada instante.

Gracias a Todos.

Resumen

Uno de los aspectos tenológicos que contribuyen a mejorar la calidad de vida de los adultos, es precisamente, el enriquecimiento de espacios físicos con sensores, equipos de video vigilancia y actuadores, que favorezcen la realización de sus actividades de la vida diaria, lo que permite descubrir patrones de acciones humanas generados a partir del movimiento y la interacción de los individuos con el ambiente, de tal manera que faciliten el monitoreo de datos y la comprensión de la actividad de los adultos mayores en entornos de vigilancia, basados en tecnología, con el propósito de detectar automáticamente patrones anormales, que afecten su salud o puedan poner en riesgo su vida. Todas estas actividades básicas les confieren a los adultos mayores la posibilidad de interactuar en comunidad con la tranquilidad de una atención médica personalizada y funcional a través de la implementación de tecnología. Aunque la lista de actividades que puede realizar una persona es extensa, este estudio se enfocó en aquellas que se desarrollan en ambientes indoor. El reconocimiento de actividades humanas es un ámbito de investigación que se suscribe a un marco investigativo, que es el estudio de las actividades de la vida diaria.

Monitorear las actividades humanas de la vida diaria es una forma de describir el estado funcional y de salud de un ser humano. El rápido crecimiento poblacional de adultos mayores ha provocado un aumento en la demanda del cuidado personal, particularmente para personas con afectaciones propias de la demencia senil, debido a la correlación que existe entre esta y el deterioro de la memoria, el intelecto, el comportamiento y la consecuente disminución de la capacidad para realizar actividades de la vida diaria.

Por tanto, surge la necesidad de realizar este proyecto, que establece un modelo predictivo de actividades de la vida diaria realizadas por habitantes en ambientes indoor, mediante el uso de técnicas de clasificación y selección basadas en Machine Learning.

Palabras clave: reconocimiento de actividades humanas, machine learning, técnicas de selección, técnicas de clasificación, actividades de la vida diaria, bases de datos

Abstract

One of the technological aspects that contribute to improving the quality of life of adults, is precisely the enrichment of physical spaces with sensors, video surveillance equipment and actuators, which favor the performance of their daily life activities, which allows discover patterns of human actions generated from the movement and interaction of individuals with the environment, in such a way that they facilitate the monitoring of data and the understanding of the activity of older adults in surveillance environments, based on technology, with the purpose of automatically detecting abnormal patterns, which affect your health or could endanger your life. All these basic activities give older adults the possibility of interacting in community with the tranquility of a personalized and functional medical attention through the implementation of technology. Although the list of activities that a person can perform is extensive, this study focused on those that take place in indoor environments. The recognition of human activities is a field of research that subscribes to an investigative framework, which is the study of activities of daily life.

Monitoring the human activities of daily life is a way of describing the functional and health status of a human being. The rapid population growth of older adults has caused an increase in the demand for personal care, particularly for people with affectations typical of senile dementia, due to the correlation that exists between this and the deterioration of memory, intellect, behavior and the consequent decrease in the ability to carry out activities of daily living.

Therefore, the need arises to carry out this project, which establishes a predictive model of activities of daily life carried out by inhabitants in indoor environments, through the use of classification and selection techniques based on Machine Learning.

KeyWords: human Activities Recognition (HAR), machine learning, selection techniques, classification techniques, Activities of Daily Life (ADL), dataset

Contenido

LISTA DE TABLAS Y FIGURAS……..……11

1. INTRODUCCIÓN.....13

1.1 CONTEXTO 13

1.2 PROBLEMÁTICA ABORDADA Y MOTIVACIÓN 14

1.3 JUSTIFICACIÓN.....16

1.4 OBJETIVOS.....17

1.4.1. Objetivo general 17

1.4.2. Objetivos específicos.....17

1.5 MAPA DEL DOCUMENTO.....18

2. FUNDAMENTACIÓN CONCEPUAL 20

2.1 RECONOCIMIENTO DE ACTIVIDADES DE LA VIDA DIARIA 20

2.2 DATASET.....23

2.3 TÉCNICAS DE INTELIGENCIA ARTIFICIAL.....25

2.3.1 Machine Learning (ML) 26

2.3.2. Tecnicas de selección de caracteristicas.....32

2.3.3. Métricas de evaluación 36

3. TRABAJOS RELACIONADOS.....41

3.1 IMPLEMENTACIÓN DE TÉCNICAS DE PREPROCESAMIENTO EN EL HAR 41

3.2 TÉCNICAS DE BASADAS EN APRENDIZAJE SUPERVISADO Y NO SUPERVISADO APLICADAS PARA EL HAR.....44

3.3 ÁMBITO DE DESARROLLO EN HAR.....46

4. METODOLOGÍA.....49

4.1 METODOLOGIAS PROPIAS DE PROCESOS BASADOS EN MINERÍA DE DATOS 49
4.1.1. Metodología KDD 49
4.1.2. Metodología SEMMA 51
4.1.3. Metodología CRIPS-DM 51
4.2 METODOLOGÍA EMPLEADA 52
4.3 MODELO FUNCIONAL PREDICTIVO 54
4.3.1. Integración y depuración 55
4.3.2. Agrupamiento de instancias y aplicación de técnicas de representación de características por subconjunto de datos 57
4.3.3. Entrenamiento y prueba de modelos para la clasificación 58
4.3.4. Evaluación de las métricas de calidad de los modelos 59
5. EXPERIMENTACIÓN 60
5.1 ESCENARIO No 1: EVALUACIÓN DEL MODELO UTILIZANDO EL DATASET PREPROCESADO 60
5.2 ESCENARIO No 2: EVALUACIÓN DEL MODELO UTILIZANDO EL DATASET SEGMENTADO 62
5.3 ESCENARIO No 3: EVALUACIÓN DEL MODELO UTILIZANDO EL DATASET SEGMENTADO Y BALANCEADO 64
5.4 ESCENARIO No 4: EVALUACIÓN DEL MODELO HIBRIDANDO TÉCNICAS DE SELECCIÓN Y CLASIFICACIÓN CON EL DATASET SEGMENTADO Y BALANCEADO 67
6. CONCLUSIONES 71

7. REFERENCIAS 74

Lista de tablas y figuras

Tablas

Pág.

Tabla 1. Actividades y número de veces realizadas en el dataset …… 23

Tabla 2. Categorización de los sensores ……. 24

Tabla No. 3 Matriz de confusión …… 36

Tabla No. 4 Comparación del enfoque propuesto en (Fahad et al., 2015) con ET-KNN, sin selección de caracteristicas y balanceo de datos, utilizando los dataset kyoto y kasteren …… 48

Tabla No 5. Comparativo de las etapas de las metodologías KDD, SEMMA Y CRISP-DM …… 53

Tabla No. 6. Descripción de las características del dataset Preprocesado …… 56

Tabla No 7. Descripción de las características del dataset segmentado …… 57

Tabla No. 8. Resultados de la evaluación de técnicas de clasificación con el dataset complete …… 61

Tabla No. 9. Resultados Técnicas de Clasificación dataset segmentado …… 62

Tabla No. 10. Comparación mejores resultados escenario 1 vs escenario 2 …… 63

Tabla No.11. Comparación dataset segmentado vs dataset segmentado y balanceado ……. 64

Tabla No. 12. Resultados Técnicas de Clasificación dataset completo …… 65

Tabla No. 13. Comparación mejores resultados escenario 1, escenario 2 y escenario 3 …… 65

Tabla No 14. Atributos del dataset …… 67

Tabla No 15. Prioridad de las características según técnica de selección de atributos ……. 67

Tabla No. 16. Resultados Hibridación Técnicas de Clasificación y selección …… 69

Tabla No. 17 Resultado comparativo entre el clasificador Classification Via Regression + Gain ratio y OneR + Gain Ratio …… 70

Tabla No. 18 Comparativo de las métricas de calidad del escenario1, escenario 2 y escenario 3 …… 71

Tabla No. 19 Priorización de características según técnica de selección de atributos …… 72

Tabla No. 20 Comparación de clase basado en Recall por cada uno de los escenarios …… 73

Figuras

Pág.

Figura No 1. Curva ROC. Tomada de (Burgueño et al., 1995). …… 39

Figura No 2. Etapas de la metodología KDD …… 50

Figura No 3. Metodología CRISP-DM …… 52

Figura No.4. Proceso de construcción del modelo funcional predictivo …… 55

Figura No.5. Clases del Dataset Balanceado ……... 58

1. Introducción

En este capitulo se describe el contexto del propósito de la investigación, posteriormente se plantea la problemática abordada que soporta este estudio, se describe sucintamente una justificación del proyecto, además de los objetivos que se pretenden alcanzar y por último se plantea el mapa del documento.

1.1 Contexto

Los adultos mayores normalmente presentan problemas con relación a su salud y al manejo de su calidad de vida. Existen muchos estudios, que han permitido fortalecer el bienestar social de las personas mayores, mediante el uso de la tecnología. Uno de los aspectos tenológicos que contribuyen a mejorar la calidad de vida de los adultos, es precisamente, el enriquecimiento de espacios físicos con sensores, equipos de video vigilancia y actuadores, que favorezcen la realización de sus actividades de la vida diaria (ADL) en ambientes *indoor*, lo que permite descubrir patrones de acciones humanas generados a partir del movimiento y la interacción de los individuos con el ambiente, de tal manera que faciliten el monitoreo de datos y la comprensión de la actividad de los adultos mayores en entornos de vigilancia, basados en tecnología, con el propósito de detectar automáticamente patrones anormales, que afecten su salud o puedan poner en riesgo su vida. Generando de esta manera, condiciones que les permitan a las personas mayores tener una vida cómoda, confortable e independientemente. Este tipo de herramientas tecnológicas permiten monitorear las acciones cotidianas que van desde el manejo básico de la higiene, al cuidado personal, la limpieza y hasta la preparación de las comidas. Todas estas actividades básicas les confieren a los adultos mayores la posibilidad de interactuar en comunidad (Nagi et al., 1965) con la tranquilidad de una atención médica personalizada y funcional a través de la implementación de tecnología.

Aunque la lista de actividades que puede realizar una persona es extensa, este estudio se enfocó en aquellas que se desarrollan en ambientes *indoor*. Donde prevalece el dominio de la capacidad cognitiva para la realización de dichas acciones. En aras de poder identificar este tipo de actividades se han desarrollo diversos estudios desde diferentes líneas de trabajo: desde la perspectiva estocástica, haciendo uso de las cadenas ocultas de Márkov (Lladó et al., 2020), desde el tratamiento de la incertidumbre a través de lógica difusa (Carlos et al., 2006), y desde la ontología web (Marcondes & Almeida Campos, 2008) a partir de los conceptos sobre los cuales se estructuran los desarrollos web.

El reconocimiento de actividades humanas (HAR), es un ámbito de investigación que se suscribe a un marco investigativo, que es el estudio de las actividades de la vida diaria. El HAR se soporta en los desarrollos tecnológicos para proporcionar a los investigadores una forma mas acertada de capturar datos, producto de las interacciones de los individuos con tales tecnologías y, por tanto, analizar esa información a través de diferentes herramientas de procesamiento de datos. Recientemente HAR ha recibido considerable atención en el campo de la robótica, debido a la recolección y almacenamiento de la información. Monitorear las actividades humanas de la vida diaria es una forma de describir el estado funcional y de salud de un ser humano. Por lo tanto, es relevante la implementación de tecnologías basadas en HAR para la atención médica personalizada.

1.2 Problemática abordada y motivación

El rápido crecimiento poblacional de adultos mayores, ha provocado un aumento en la demanda del cuidado personal, particularmente para personas con afectaciones propias de la demencia senil, debido a la correlación que existe entre esta y el deterioro de la memoria, el intelecto, el comportamiento y la consecuente disminución de la capacidad para realizar actividades de la vida diaria de las personas que sufren de estos trastornos (Amiribesheli et al.,

2015). Una solución relativamente eficiente para disminuir los costos concernientes al cuidado de las personas con demencia senil es complementar la dinámica de la atención formal en hospitales y hogares geriátricos con la atención informal en hogares privados mediante tecnología tipo *Smart Home* (SH), la cual tiene como objetivo ayudar a las personas a acceder a una mejor calidad de vida y garantizar que el adulto mayor pueda vivir cómoda, confortable e independientemente.

La tecnología SH se considera una forma de reducir los costos de vida y cuidado, para mejorar las condiciones sociales de las personas con necesidades especiales de atención, debido a las dificultades de movilidad y deterioro cognitivo (Ding et al., 2011). La mencionada tecnología, se ha aplicado en muchos ámbitos para ahorro de energía (Khalifa et al., 2018), seguridad y protección (Khalifa et al., 2018), detección de caídas (Cardoso & Moreira, 2016), control de la luz (Jiang et al., 2017), detección de humo y fuego (Islam, 2018), entre otros, mediante diversas soluciones como: monitoreo de video (Jalal et al., 2017), alarmas (He et al., 2012) y planificadores inteligentes (Tabuenca Dopico et al., 1993) que hacen uso de calendarios y recordatorios. Todas estas soluciones y muchas otras, habitualmente utilizan diferentes equipos para el proceso de captación de datos y generación de respuestas físicas, tales como: sensores, equipos de video vigilancia y actuadores.

El objetivo del HAR, es descubrir patrones de actividades humanas mediante el análisis de los datos generados a partir del movimiento y las interacciones de los individuos con el ambiente. HAR tiene varias aplicaciones potenciales, para facilitar el monitoreo de los datos y comprender la actividad humana en entornos de vigilancia basados en tecnología (Y. Chen & Shen, 2017), con el propósito de detectar automáticamente patrones anormales en el desarrollo de actividades.

Por tanto, surge la necesidad de realizar este proyecto, que establece un modelo predictivo de actividades de la vida diaria realizadas por habitantes en ambientes *indoor*, mediante el uso de técnicas de selección y clasificación basadas en *Machine Learning (*ML*)*. Teniendo en cuenta lo anteriormente mencionado, surge el siguiente interrogante.

¿Cómo identificar actividades de la vida diaria en ambientes *indoor* usando la inteligencia artificial?

1.3 Justificación

En la actualidad, el planeta cuenta con una población superior a 7.5 billones de personas, de las cuales, según la Organización Mundial de la Salud (OMS) (*Envejecimiento y Salud*, 2018) entre el 2015 y 2050, la proporción de residentes mayores de 60 años en el planeta casi se duplicará, pasando del 12% al 22%. El aumento de la esperanza de vida brinda oportunidades no solo a los ancianos y sus familias, también lo hace a toda la sociedad. Durante estos años extra, se pueden realizar nuevas actividades, como seguir aprendiendo, iniciar una nueva carrera o volver a una antigua afición. Sin embargo, el alcance de estas oportunidades y contribuciones depende en gran medida del factor salud.

La vejez también se caracteriza por la aparición de varios estados de salud complejos, que generalmente afloran en las etapas finales de la vida. Estas condiciones de salud se conocen comúnmente como síndrome de la vejez. Por lo general, son el resultado de múltiples factores subyacentes, que incluyen, entre otros, los siguientes: fragilidad, incontinencia urinaria, caídas, delirios y úlceras por decúbito. Estas dificultades de salud han llevado a muchas personas mayores a permanecer en casa bajo el cuidado y la supervisión de sus seres queridos, quienes velan por ofrecerles el mayor bienestar en su vida cotidiana. Sin embargo, si bien pueden ayudar a planificar la rutina de la persona mayor, no existe ningún tipo de alerta para el caso de que

estas actividades planeadas no se realicen de forma correcta, lo cual podría indicar la presencia de un posible problema que requiere de atención inmediata.

Lo anterior, permite precisar que tanto los adultos mayores que presentan un estado de salud deficiente, como aquellos cuyas condiciones personales los motivan a seguir activos y productivos, tanto en casa como en hogares geriátricos, necesitan de la combinación de métodos avanzados de seguimiento visual, optimización, reconocimiento de patrones y aprendizaje, que les provean entornos seguros y confortables y que a la vez sirvan de herramienta para facilitar el trabajo de familiares y trabajadores. Cabe resaltar que con ello también se busca recrear una tecnología que les brinde autonomía a estos adultos en ambientes *indoor*.

Es por ello que la presente investigación se enfoca en generar un modelo de predicción de actividades de la vida diaria mediante técnicas de clasificación y selección de características, con el fin de aportar en el desarrollo en esta área del conocimiento, especialmente en el ámbito de la salud, para llevar un monitoreo preciso de las actividades de los adultos mayores o personas con algún tipo de discapacidad. Los desarrollos tecnológicos permiten analizar de manera predictiva las actividades de la vida diaria contribuyendo a la identificación de patrones de manera previa, para efectos de tomar acciones en pro de la mejora de la calidad de vida del adulto mayor.

1.4 Objetivos

1.4.1. Objetivo general

Desarrollar un modelo predictivo para la identificación actividades de la vida diaria (ADL) en ambientes *indoor*, usando técnicas de clasificación basadas en *Machine Learning*.

1.4.2. Objetivos específicos

Para la consecución del objetivo general anteriormente mencionado, se han planteado los siguientes tres objetivos específicos, en coherencia con la ejecución de las fases de documentación, desarrollo y evaluación.

- Documentar referentes teóricos y prácticos referidos a conjuntos de datos utilizados para el reconocimiento de las actividades humanas en ambientes *indoor*, el uso de técnicas de preprocesamiento (selección de características) y el entrenamiento de modelos predictivos basados en *machine learning* para la clasificación.
- Desarrollar un modelo que permita predecir actividades de la vida diaria realizadas por habitantes de ambientes *indoor*, basado en la hibridación de técnicas de selección y clasificación.
- Evaluar el modelo predictivo propuesto a partir de la implementación de diferentes escenarios, validados mediante el análisis de las métricas de calidad empleadas.

1.5 Mapa del documento

El presente documento de investigación está constituido por seis (6) capítulos. Cada uno de los cuales se describe a continuación.

El primer capítulo contiene una introducción acerca del reconocimiento de las actividades de la vida diaria y la importancia de descubrir patrones de actividad física, mediante el análisis de los datos de movimiento capturados a través de múltiples sensores. Adicionalmente se incluyen los objetivos que inicialmente fueron planteados y el mapa del documento.

En el segundo capítulo se abordan los ejes temáticos que fundamentan la investigación, los cuales son: Reconocimiento de Actividades de la Vida Diaria - ADL, *datasets* y técnicas de Inteligencia Artificial, específicamente las basadas en *Machine Learning*. Con respecto al primero, su objetivo es reconocer las actividades de la vida cotidiana, a través de diferentes dispositivos electrónicos, los cuales permitirán almacenar grandes cantidades de datos (conjuntos de datos), que luego serán usados para implementar modelos predictivos del comportamiento humano. Con respecto al segundo eje, se describen en detalle la recopilación de información y los métodos que se pueden utilizar para construir estos datos. Finalmente, el tercer eje detalla

diferentes técnicas de selección y clasificación de características mediante el aprendizaje automático, de manera que los indicadores de las métricas de calidad faciliten el proceso de comparación y análisis de los mejores resultados, con el propósito de identificar el modelo predictivo mas óptimo.

En el tercer capítulo se realiza un análisis de la revisión de la literatura y producto de esto se han identificado una serie de referentes valiosos en correlación con la implementación de técnicas de preprocesamiento en el HAR, técnicas basadas en aprendizaje supervisado y no supervisado y ámbitos de desarrollo en el HAR.

En el cuarto capítulo se presentan diferentes metodologías relacionadas con el proceso de Minería de Datos, en aras de sustentar la metodología aquí utilizada. Con base en ella se muestra un diagrama del modelo funcional predictivo propuesto con su respectiva explicación y por último se enumeran los diferentes escenarios de experimentación que se han recreado para el desarrollo de esta investigación.

En el quinto capítulo se han recreado una serie de escenarios de experimentación los cuales reflejan la evaluación del modelo con el *dataset* preprocesado, evaluación del modelo utilizando el *dataset* segmentado, evaluación del modelo utilizando el *dataset* segmentado y balanceado, evaluación del modelo hibridando técnicas de selección de características y clasificación con el *dataset* segmentado.

En el sexto capítulo refleja las conclusiones a las que se ha llegado al finalizar el modelo predictivo de este trabajo de investigación, en las que se dan a conocer los resultados obtenidos, a su vez, se proponen trabajos futuros.

2. Fundamentación concepual

El presente estudio se fundamenta a partir de los ejes conceptuales relacionados con el reconocimiento de actividades de la vida diaria, *dataset* y técnicas de inteligencia artificial. En este capítulo se hace un análisis detallado de cada uno de estos ejes. En cuanto al primero, se conceptualiza en el reconocimiento de las actividades de la vida diaria mediante diferentes dispositivos capaces de capturar los movimientos y el tiempo de duración de una actividad en ambientes *indoor*; además de la importancia que aporta a la sociedad el poder realizar entrenamiento a partir de grandes volúmenes de datos (*dataset*), con el propósito de implementar modelos predictivos del comportamiento humano. En cuanto al segundo eje, se detalla la colección de datos y las metodologías que permiten la construcción de estos. Finalmente, en el tercer eje se detallan las diferentes técnicas de selección de características y de clasificación usando ML, permitiendo así, realizar un análisis comparativo de las métricas de calidad de dichas técnicas, buscando generar un modelo predictivo con los mejores resultados obtenidos.

2.1 Reconocimiento de actividades de la vida diaria

En los últimos años, con el desarrollo y la integración de la tecnología electrónica, la mayoría de los teléfonos inteligentes se han generalizado en todo el mundo y se han convertido en dispositivos indispensables en nuestra vida diaria. El proceso de mejora continua de tecnologías de detección como acelerómetros, giroscopios, GPS, magnetómetros y termómetros integrados en estos teléfonos inteligentes han permitido la detección del contexto consciente del usuario y proporcionado un servicio más personalizado. A través de abundantes sensores en teléfonos inteligentes, es posible captar señales de secuencia de alta frecuencia y precisión en tiempo real.

Según Yuan (Yuan et al., 2019) los datos capturados en bruto construyen un puente entre los sensores y las actividades humanas, que lleva a una rápida popularidad del HAR a partir de

los datos detectados. El HAR desempeña un papel importante en aras de ofrecer mayor comodidad para nuestras vidas, en muchos campos de aplicación, como: el monitoreo de la salud (Nazabal, Garcia-Moreno, Artes-Rodriguez, & Ghahramani, 2016), el monitoreo deportivo (Eskaf, Aly, & Aly, 2016), la detección de caídas (Cardoso & Moreira, 2016), la seguridad social (Vinayak et al., 2018) y la implementación de tecnologías en hogares inteligentes (Raeiszadeh & Tahayori, 2018), entre otros. Los *Smartphone* actuales poseen varios beneficios en relación con el proceso de reconocimiento de actividades humanas. Primero, tal como se plantea en (Yuan et al., 2019), son dispositivos de alta disponibilidad, es decir, casi todas las personas tienen acceso a un teléfono inteligente hoy en día y para el HAR con este tipo de dispositivos es suficiente. En segundo lugar, muchos sensores integrados en el teléfono inteligente hacen que pueda recopilar datos de actividades desde una perspectiva diferente. En tercer lugar, la capacidad de cálculo de alto rendimiento del teléfono inteligente avanza en línea y en tiempo real en el reconocimiento de actividades complejas. En cuarto lugar, los componentes de comunicación de la naturaleza, como GPS, *Bluetooth* y *WiFi*, hacen que no solo pueda reconocer la microactividad, sino que también puede descubrir la macroactividad, que proporciona una aplicación más amplia para *smartphones*.

El problema de la detección en interiores y exteriores ha sido considerado por muchos investigadores y puede clasificarse en dos enfoques principales uno basado en la visión y el otro en el sensor. Los métodos basados en la visión dependen del *hardware* utilizado y de los parámetros condicionales y no pueden ser una solución de diseño óptimo para detectar el entorno utilizando una plataforma de dispositivo móvil inteligente. Por lo tanto, el enfoque se ha movido hacia un punto de vista basado en sensores para encontrar una solución deseable para plataformas de dispositivos móviles. En (Walter et al., 2013) se introdujo un nuevo método que

incluye la gravedad, la luz ambiental y los campos magnéticos para detectar el ambiente. Grove (Groves, Martin, Voutsis, Walter, & Wang, 2013) distinguió el contexto ambiental midiendo el número de puntos de acceso a la estación de señal vecina GSM (GNSS) y *Wi-Fi*. Los resultados de su estudio mostraron que GNSS no podía usar mediciones para distinguir los lugares interiores de los exteriores. Ravindranath (Sivalingam, 2010) introdujo la capacidad del estado de bloqueo del GPS para inferir las condiciones del ambiente indirectamente. Sin embargo, a pesar de la precisión del GPS en muchas aplicaciones, se puede agotar la batería completamente después de aproximadamente seis (6) horas de uso continuo, y solo es precisa en entornos de espacio abierto. (W. Wang et al., 2016). Debido a los inconvenientes anteriores de la solución basada en GPS, considerando que los sensores integrados de bajo costo en dispositivos móviles inteligentes proporcionan solo algunos recursos para el problema de detección en interiores y en exteriores, se emplean dispositivos inteligentes dotados de sensores integrados como el giroscopio, el micrófono, la luz, el magnetómetro y el acelerómetro para supervisar el movimiento del celular, como la inclinación, la vibración, la rotación o el balanceo, movimiento que suele ser un reflejo de la interacción directa del usuario con el entorno físico en el que se encuentra el dispositivo, de tal manera que se pueda aprovechar y manipular la tecnología para detectar posibles problemas de salud, físicos e inclusive cognitivos de manera que los adultos mayores puedan ser asistidos con prontitud evitando que su situación vital empeore. Estas tecnologías inteligentes se encuentran integradas por componentes de *hardware* (basados en sensores, en IoT, etc) y software, que son la clave para el desarrollo de aplicaciones y servicios para la interconexión de personas y cosas necesarias para solucionar problemas de la vida cotidiana.

Una tendencia en investigación es el uso de sistemas HAR basados en técnicas de IA (Inteligencia artificial); tales sistemas requieren un proceso de entrenamiento a partir de grandes volúmenes de datos (*Dataset*), con el propósito de implementar modelos predictivos del comportamiento humano.

Dichos sistemas de almacenamiento de datos se convierten en pieza fundamental para la revolución del procesamiento de estos, por lo que se hace necesario profundizar en el concepto y la influencia del *dataset* en este proyecto de investigación.

2.2 Dataset

Un *dataset* es una colección de datos que pueden provenir de fuentes heterogéneas y que mediante técnicas de preprocesamiento son tabulados o estructurados, con la finalidad de analizarlos para identificar relaciones no triviales entre ellos.

En la última década, el aprendizaje automático y las tecnologías de computación dominantes han madurado hasta el punto en que no solo se integran con nuestras vidas, también brindan un soporte automatizado y sensible al contexto en nuestros entornos cotidianos. Una aplicación física de tal sistema son las casas inteligentes. En el entorno doméstico, el software informático, desempeña el papel de un agente inteligente que percibe el estado del entorno físico y los residentes que interactúan con sensores.

Gracias a los procesos de captación de datos, los sensores integrados en el hogar permiten la lectura de las interacciones de los individuos con dichos sensores, mientras que los residentes realizan sus rutinas diarias. Las lecturas de los sensores son recopiladas gracias a la conexión de dispositivos en el ambiente *indoor*, mediante tecnologías y protocolos de comunicación. Los datos recolectados son almacenados en bases de datos que modelos pre-entrenados utilizan para generar conocimiento útil, a partir del análisis de patrones, la generación de predicciones y la identificación de tendencias de comportamiento.

El proyecto casas de la universidad de WSU (Washington State University) (Cook, Crandall, Thomas, & Krishnan, 2013), presenta un ambiente basado en el uso de técnicas de IA, donde el estado de los residentes y su entorno físico se perciben mediante sensores de manera que se pueda mejorar la comodidad, la seguridad y la productividad de los residentes. El conjunto de datos de referencia, adoptado en los escenarios de experimentación recreados en el marco de este proyecto, es el de actividades de la vida diaria (ADL *Activities*) de CASAS KYOTO, el cual contiene un sinnúmero de datos que representan eventos de sensores que detectan el movimiento, en un ambiente *indoor*, recreado por la WSU. Para generar el *dataset* de ADL *Activities*, los investigadores del laboratorio CASAS reclutaron a veinte (20) voluntarios participantes, para realizar cinco (5) actividades, las cuales se describen en la siguiente tabla:

Tabla 1.

Actividades y número de veces realizadas en el dataset

Actividad	Total
Realizar llamada telefónica	24
Lavarse las manos	24
Cocinar	24
Comer	24
Limpiar	24

Para la recolección de los datos se usaron diferentes sensores, que son referenciados con un identificador, ver tabla 2.

Tabla 2

Categorización de los sensores

Sensor	Categorización
Sensores de movimiento	M01 – M26

Sensores para la utilización de elementos de cocina	I01 – I05
Sensor en el contenedor de las medicinas	I06
Sensor en utensilios de cocina	I07
Directorio telefónico	I08
Sensor de gabinetes	D01
Sensor de agua	AD1-A - AD1-B
Sensor de encendido de la cocina	AD1-C
Uso del teléfono	Asterisk

Fuente propia del autor

El sitio web oficial de CASAS contiene una amplia variedad de conjuntos de datos y herramientas. Para cada uno de estos se detalla lo siguiente: el nombre del banco de pruebas, el número de residentes o participantes ya sea anotado o no. Además, los archivos de cada *dataset* están disponibles para descargar. Según Crandall (Crandall & Cook, 2013) el *dataset* más referenciado del repositorio de CASAS es *Daily life* 2010-2012 Kyoto, también llamado ADL *Activities*.

En la actualidad existe una gran variedad de *dataset,* los cuales pueden ser utilizados dependiendo de la necesidad del usuario final. Para llevar a cabo este proyecto se usarán una serie de técnicas de inteligencia artificial, las cuales permitirán establecer un modelo predictivo aplicando técnicas de selección de características y de clasificación, basadas en ML, permitiendo así realizar un análisis comparativo de las métricas de calidad de dichas técnicas, buscando generar un modelo predictivo con los mejores resultados obtenidos.

2.3 Técnicas de inteligencia artificial

Las técnicas basadas en IA han incluido un sinnúmero de aplicaciones dentro del campo de la Ingeniería. Estas van desde la automatización de importantes procedimientos en la Industria

y las empresas, hasta el campo del Control de Procesos. La tecnología tipo *Smart Home* (SH), están diseñadas para ayudar a los residentes de las casas a mejorar sus actividades de la vida diaria y por lo tanto enriquecer la calidad de vida preservando su privacidad. Un sistema SH suele ser equipado con una colección de *software* interrelacionado con componentes de *hardware* para monitorear el espacio vital mediante la captura del comportamiento del residente y sus ocupaciones. Al hacerlo, el sistema puede informar sobre riesgos, situaciones y tomar medidas en nombre del residente a su satisfacción.

Actualmente, la IA se está aplicando en las actividades relacionadas con los seres humanos, para el análisis y procesamiento de datos usando técnicas de *ML*.

2.3.1 Machine Learning (ML).

ML es una disciplina científica del campo de la IA, que se encarga de crear sistemas que aprenden automáticamente; esto quiere decir que identifica patrones complejos en millones de datos, mediante algoritmos que revisan la información y predicen comportamientos futuros.

En nuestro mundo cada vez más complejo, el campo de la analítica ha aumentado dramáticamente su importancia. La intuición ya no es suficiente en la toma de decisiones, pero esta debe combinarse con el apoyo de la enorme cantidad de datos disponibles en la actualidad. Para Kaj-Mikael Björk (Björk et al., 2016) los problemas surgen en al menos dos situaciones: cuando los datos son imprecisos por naturaleza y cuando los datos están incompletos. Ambas situaciones son problemáticas y deben abordarse de manera adecuada. "*A new application of machine learning in health care*" apunta a combinar la investigación básica y la aplicada, en ciertos contextos, obteniendo nuevos métodos de manipulación de datos imprecisos e incompletos basados en redes neuronales de aprendizaje automático.

Según (LeCun, Bengio, & Hinton, 2015), el aprendizaje profundo (*Deep learning*) permite que los modelos computacionales, compuestos por múltiples capas de procesamiento, aprendan representaciones de datos con diferentes niveles de abstracción. En los últimos años las redes neuronales artificiales han sido ampliamente utilizadas en investigaciones que buscan efectuar reconocimiento de patrones y procesos de aprendizaje automático; en Schmidhuber (Schmidhuber, 2015) se hace una revisión del aprendizaje profundo supervisado, además del aprendizaje no supervisado y el cálculo evolutivo.

Se pueden utilizar una variedad de métodos de ingeniería de conocimiento y procesamiento de datos para analizar la información, los mas referenciados métodos son: árboles de decisión (Amiribesheli et al., 2015), máquinas de soporte vectorial - SVM (Fleury et al., 2010), clasificadores Naive Bayesian (Andrew McCallum, 1998), modelos ocultos de Markov (Eddy, 1998), clasificadores basados en lógica difusa (Amiribesheli et al., 2015), redes neuronales artificiales – ANN (Murata et al., 1994), tablas de decisiones (Du & Hu, 2014), y árboles de modelos logísticos - LMT (W. Chen et al., 2017). A continuación, una descripción mas detallada de algunos de estos.

Árboles de decisión (Amiribesheli et al., 2015): se utiliza un árbol de decisión (*Decision Tree* - DT) para modelar la relación entre datos de entrada y la salida correspondiente. Un DT puede usarse para cualquier clasificador. Teniendo en cuenta que, si la salida es discreta, implica etiquetas de clase y si la salida es continua implica etiquetas de regresión. Un árbol de decisión consta de nodos que representan entidades y ramas que son los valores de las características. Los nodos de las hojas representan las etiquetas de clase. El modelo de clasificación basado en DT es uno de los más utilizado y conocido, debido a su facil comprensión.

Según (Ricardo et al., 2015), la calidad del árbol de decisión depende del tamaño y la precisión de la clasificación. Se escoge un subconjunto del conjunto del *dataset* (entrenamiento) y se crea un árbol de decisión. Si no arroja la respuesta para los objetos del conjunto de prueba, una selección de excepciones se agrega al conjunto del *dataset,* continuando el proceso hasta encontrar el conjunto de decisiones correctas.

Los algoritmos de clasificación mas utilizados, de la categoría árboles de decisión, son: id-3, c4.5, CART, *Sprint* y j48 (Ricardo et al., 2015).

Máquinas de soporte Vectorial (Fleury et al., 2010): Las máquinas de soporte vectorial (*Support Vectorial Machine* - SVM), son métodos de clasificación bastante populares y se han utilizado en varias aplicaciones, como la identificación de caras, la categorización de textos y la clasificación de *stocks*. Una SVM construye un hiperplano óptimo en forma de superficie de decisión, de modo que el margen de separación entre las dos clases en los datos se amplía al máximo.

En las SVM las magnitudes de entrada son mapeados no linealmente a un espacio de características de muy alta dimensión. En este espacio de características se construye una superficie de decisión línea (Cortes et al., 1995).

Según (Han et al., 2012), los SMV utilizan un mapeo no lineal para transformar los datos originales de la formación, en una dimensión superior. Dentro de esta nueva dimensión, busca el hiperplano de separación óptimo lineal (es decir, un "límite de decisión" que separa las tuplas de una clase de otra). Los SVMs pueden ser usados para la predicción numérica, así como para la clasificación. Se han aplicado a una serie de áreas, incluyendo el reconocimiento manuscrito de dígitos, el reconocimiento de objetos y la identificación de hablantes, así como pruebas de predicción de series de tiempo de referencia.

Clasificadores Naive Bayesian (Andrew McCallum, 1998): Este clasificador es el más simple, ya que asume que todos los atributos de los ejemplos son independientes entre sí, dado el contexto de la clase, a esta se le llama “Asunción Bayesiana". Si bien esta suposición es claramente falsa en la mayoría de las tareas del mundo real, este a menudo realiza una clasificación adecuada. Esta paradoja se explica por el hecho de que la estimación de la clasificación es solo una función del signo (en casos binarios) de la estimación de la función; la aproximación de la función aún puede ser deficiente, mientras que la precisión de la clasificación sigue siendo alta según (P. Domingos and M. pazzani, 1997). Debido a la suposición de independencia, los parámetros para cada atributo se pueden aprender por separado, y esto simplifica enormemente el aprendizaje, especialmente cuando el número de atributos es grande.

La clasificación de documentos es solo un dominio con una gran cantidad de atributos. Los atributos de los ejemplos que deben clasificarse son palabras, y el número de palabras diferentes puede ser bastante grande. Mientras que algunas tareas simples de clasificación de documentos se pueden realizar con precisión como el tamaño de vocabulario menor a cien, además de muchas tareas complejas sobre datos del mundo real.

Modelos ocultos de Markov (HMM) (Eddy, 1998): Es útil comprender la generalidad y la simplicidad relativa de la teoría de HMM. Un HMM describe una distribución de probabilidad sobre un número potencialmente infinito de secuencias. Debido a que una distribución de probabilidad debe sumar a uno, los "puntajes" que un HMM asigna a las secuencias están restringidos. La probabilidad de una secuencia no puede aumentarse sin disminuir la probabilidad de una o más secuencias. Es esta restricción fundamental del modelado probabilístico es la que permite que los parámetros en un HMM tengan óptimos no triviales.

Clasificadores basados en lógica difusa (Amiribesheli et al., 2015): Como extensión de la teoría de conjuntos clásica, Zadeh en 1965 define un conjunto difuso como "... una clase de objetos con un continuo grado de afiliación. Tal conjunto se caracteriza por una función de pertenencia (característica) que asigna a cada objeto un grado que oscila entre cero y uno". Los conjuntos difusos se han utilizado para desarrollar la lógica difusa. Muchos estudios de SH han aplicado lógica difusa para construir Sistemas de seguimiento y predicción.

Redes Neuronales Artificiales - ANN (Murata et al., 1994): Es un modelo informático compuesto por una serie de neuronas simples, altamente interconectadas con elementos de procesamiento.

Los elementos de procesamiento fundamentales de una ANN son las neuronas artificiales (o nodos) que están interconectadas por enlaces ponderados que forman capas. Normalmente en una ANN hay una capa de entrada y una capa de salida y un número de capas ocultas que varía dependiendo de la complejidad del problema en cuestión. Las neuronas transforman la entrada ponderada en salida, utilizando una función de activación que puede tomar diferentes formas lineales y no lineales. El proceso por el cual los pesos son ajustado se llama aprendizaje. Un número de ANNs no lineal se sabe que realiza aproximadores de función.

Existen varios parámetros que definen la arquitectura de una red neuronal: el tipo de conexión, regla de aprendizaje y funciones de activación. Debido a estos parámetros de conformación, hay diferentes tipos de ANN por ejemplo: Perceptrón de múltiples capas - MLP (Gaikwad et al., 2019), *Echo State Networks* - ESN (L. Wang, Wang, & Liu, 2016), redes de función de base radial - RBFN (G.-B. Huang, 2015), máquina de Boltzmann (Liu et al., 2017).

Tablas de decisiones (Du & Hu, 2014)**:** Una tabla de decisión también puede verse como una serie de reglas de decisión. El conjunto de reglas de decisión inducidas por las

aproximaciones definidas usando relaciones de dominancia, dan en general, una representación más sintética del conocimiento contenido en la tabla de decisiones que el conjunto de reglas inducidas a partir de aproximaciones clásicas definidas usando relaciones de indiscernibilidad y son más comprensibles y aplicables para los usuarios debido a la sintaxis más general de las reglas.

Existen al menos cuatro fuentes de inconsistencias en las tablas de decisión, que se enumeran a continuación: 1) vacilación en la evaluación de los valores de los atributos de decisión, 2) errores en el registro, medición y observación, 3) atributos de condición faltantes relacionados con la evaluación de los valores de los atributos de decisión, 4) la naturaleza inestable del sistema representado por la tabla de decisiones y similares. Estas inconsistencias no pueden considerarse como un simple error o ruido. Para adquirir reglas, de tablas de decisión inconsistentes, se necesitan reducciones relativas de atributos. Skowron y Rauszer introdujeron el método de matriz de discernibilidad que se convirtió en un enfoque popular para enumerar todas las reducciones en el Rough set (Yiyu, 2007).

Árbol basado en modelo logístico - LMT (W. Chen et al., 2017): El LMT es un modelo de clasificación, que combina métodos de aprendizaje basados en árboles de decisión y regresión logística - LR (Salzberg, 1994). En la variante logística la ganancia de información se usa para dividir. El algoritmo *LogitBoost* (Landwehr et al., 2005) se utiliza para producir un modelo de LR en cada nodo en el árbol, y el árbol se poda usando un algoritmo CART (Breiman et al., 2017). El LMT utiliza validación cruzada para encontrar una serie de iteraciones, para evitar el sobreajuste de datos de entrenamiento. El método de regresión logística lineal se usa para calcular las probabilidades posteriores de los nodos foliares en el modelo LMT.

El proceso de clasificación puede mejorar si se utilizan técnicas de selección de características, estas permiten asignar la priorización o la relevancia de los atributos por medio del criterio de clase, obteniendo así, una estructura de atributos que incidan directamente sobre el modelo y que a su vez tengan mayor relevancia con respecto a la clasificación.

2.3.2. Tecnicas de selección de características.

Algunos de los atributos contenidos en el *dataset* son irrelevantes, por eso es necesario eliminarlos, además algunos algoritmos de minería no funcionan adecuadamente con grandes cantidades de características o atributos. Por lo tanto, es necesario aplicar técnicas de selección de características antes de que se aplique cualquier tipo de algoritmo de minería. El objetivo principal de la selección de características es evitar el *overfitting* y mejorar el rendimiento del modelo además de proporcionar modelos más rápidos y rentables. La selección de características agrega una capa extra de complejidad en el modelado. Según (Daelemans et al., 2003), no solo es encontrar parámetros óptimos para un conjunto completo de características, se debe encontrar el primer subconjunto de características óptimas y los parámetros del modelo deben optimizarse. Algunas de las técnicas más utilizadas para clasificación son: *Gain Ratio* (Karegowda et al., 2010), *Info Gain* (Dietterich, 1997), *OneR* (Kumar et al., 2013), *Symmetrical Uncert* (Parimala & Nallaswamy, 2011), *RELIEF* (Kira & Rendell, 1992) y *Chi Squared* (Bidgoli & Naseriparsa, 2018) las cuales se describen a continuación.

Gain Ratio

Según (Karegowda et al., 2010), esta técnica pertenece a la categoría árboles de decisión, estos son una estructura simple donde los nodos no terminales representan pruebas en uno o más atributos y los nodos terminales reflejan los resultados de las decisiones. *Gain Ratio* es una modificación de la *Information gain*, que tiene en cuenta la cantidad y el tamaño de los nodos

secundarios en los que un atributo divide el conjunto de datos con respecto a la clase. Esto atenúa la preferencia que tiene el método de obtención de información por atributos con un gran número de valores posibles. Ver (1).

$$Gain\ ratio = \frac{H(y) + H(x) - H(y,x)}{H(x)} \quad (1)$$

Info Gain

Esta técnica, definida en (Dietterich, 1997), consiste en calcular la información mutua (también llamada *information gain*) entre cada característica de entrada y la clase. La información mutua entre dos variables aleatorias es la reducción media de la incertidumbre sobre la segunda variable, dado un valor de la primera. Para la característica discreta j, el peso de la información mutua w_j se puede calcular como (2):

$$w_j = \sum_v \sum_c P(y = c, x_j = v) * log \frac{P(y = c, x_j = v)}{P(y = c)P(x_j = v)} \quad (2)$$

donde P (y = c) es la proporción de ejemplos de entrenamiento en la clase c, y P (xj = v) es la probabilidad de esa característica, j adquiere valor v. Para valores reales de características, las sumas se convierten en integrales que deben ser aproximadas.

Un problema con la ponderación de *information gain* es que trata cada característica de forma independiente. Para las características cuyo poder predictivo solo es aparente en combinación con otras características, la información mutua asigna un peso de cero. Por ejemplo, una clase difícil, con problemas de aprendizaje implica el aprendizaje de funciones de paridad con características aleatorias irrelevantes. Una función de paridad sobre una característica binaria es igual a uno (1) si y solo si, un el número impar de características es igual a uno (1).

Supongamos que definimos un problema de aprendizaje en el que hay cuatro (4) características relevantes y 10 características binarias irrelevantes, y la clase es la paridad de las cuatro (4) características relevantes. Los pesos de información mutua de todas las características serán aproximadamente cero utilizando la fórmula anterior.

OneR

OneR, definida en (Kumar et al., 2013), es la abreviatura de "*One Rule*". Es un algoritmo que genera un árbol de decisión de un nivel. Éste es capaz de inferir una clasificación típicamente simple, pero precisa reglas de un conjunto de instancias. El algoritmo *OneR* crea una regla para cada atributo en los datos de entrenamiento, y luego selecciona la regla con la tasa de error más pequeña como su regla. Para crear una regla para un atributo, debe determinarse la clase más frecuente para cada valor de atributo. La clase más frecuente es simplemente la clase que aparece más a menudo para ese valor de atributo. Una regla es simplemente un conjunto de atributos de valores ligados a su clase mayoritaria.

Symmetrical Uncert

Symmetrical Uncert, definida en (Parimala & Nallaswamy, 2011)**,** es otro método que se ideó para compensar el sesgo de la ganancia de información hacia las características con más valores. Aprovecha la propiedad simétrica de la ganancia de información. La incertidumbre simétrica entre las características y el concepto de destino se puede utilizar para evaluar la bondad de las características para la clasificación, ver (3).

$$Symmetrical\ uncertainty\ = 2\ \frac{Gain}{H_{(y)}\ +\ H_{(x)}} \qquad (3)$$

Relief

Relief, definido en (Kira & Rendell, 1992)**,** es un algoritmo basado en el peso de las características inspirado en aprendizaje basado en instancias. Dados los datos de entrenamiento S, tamaño de muestra m, y un umbral de relevancia dos (2), *Relief* detecta aquellas características que son estadísticamente relevantes para el concepto de destino y codifica un umbral de relevancia. Supongamos que la escala de cada característica es nominal (incluido booleano) o numérico (entero o real). Según Kira (Kira & Rendell, 1992) la diferencias de valores de características entre dos instancias X y Y están definidos por la siguiente función diff. Cuando xk y yk son nominales, ver (4).

$$diff_{(x_k,y_k)} = \begin{cases} 0 \ < \ si\ x_k\ y\ y_k\ son\ iguales \ > \\ 1 < si\ x_k\ y\ y_k\ son\ diferentes > \end{cases} \quad (4)$$

Cuando xk y yk son númericos. Ver (5).

$$Diff(x_k, y_k) = (x_k - \ y_k)/nu_k \quad (5)$$

donde nuk es una unidad de normalización para regular los valores de diff en el intervalo [0, l].

Chi Squared (Bidgoli & Naseriparsa, 2018)

Chi-Squared justiprecia el valor de una característica calculando la cuantía de la estadística con respecto a la clase. La hipótesis inicial H0 es el supuesto de que las dos características son no relacionadas, y se prueba mediante la ecuación *Chi-Squared* como se muestra en la formula (6).

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - \ E_{ij})^2}{E_{ij}} \quad (6)$$

Donde Oij es la frecuencia observada y Eij es la frecuencia esperada, afirmada por la hipótesis nula. Los valores mayores de X^2, es la mayor evidencia contra la hipótesis H0.

Finalmente, el reconocimiento de actividades humanas permite identificar los patrones de comportamiento de los individuos productos de sus interacciones con diferentes dispositivos o sensores. Gracias a la captación de datos, estos son almacenados y tabulados en *dataset*, que, posteriormente usando técnicas de selección de características, permitirán realizar un análisis comparativo de las métricas de calidad de dichas técnicas, buscando generar un modelo predictivo con los mejores resultados obtenidos.

2.3.3. Métricas de evaluación.

A medida que se realiza el entrenamiento de un modelo predictivo se debe evaluar que tan bueno es éste. Para dicho propósito, la "matriz de confusión" es utilizada en procesos de aprendizaje automático, con el propósito de comparar información entre las categorizaciones reales y las predicciones efectuadas por el sistema de clasificación. Según (Deng et al., 2016), una matriz de confusión tiene dos dimensiones, una dimensión está indexada por el valor real de las clases de un objeto, y la otra, está por indexada por las clases a prededir por el clasificador. La matriz está compuesta por métricas básicas de calidad, son: Falsos Positivos (FP), Falsos Negativos (FN), Verdaderos Positivos (VP) y Verdaderos Negativos (VN). Dichas métricas hacen parte de la matriz de confusión, como se muestra en la Tabla No 3. A continuación, se describe sucintamente cada una de las métricas básicas.

Verdaderos Positivos (TP): Estos se refieren a las tuplas positivas que fueron correctamente etiquetadas por el clasificador.

Verdaderos Negativos (TN): Estas son las tuplas negativas que fueron correctamente etiquetadas por el clasificador.

Falsos Positivos (FP): Estas son las tuplas negativas que fueron etiquetadas incorrectamente como positivas por el clasificador.

Falsos Negativos (FN): Estas son las tuplas positivas que fueron mal etiquetadas como negativas por el clasificador.

Tabla 3

Matriz de confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuente: (Xu et al., 2020)

Para evaluar las diferentes técnicas, a partir de las cuales se construyen los modelos predictivos para procesos de clasificación, se utilizan las métricas de calidad que se detallan a continuación.

Sensibilidad:

La métrica Sensibilidad es "la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo, respecto a la condición que estudia la prueba. Razón por la que también es denominada tasa de verdaderos positivos (TPR)", según (López de Ullibarri & Píta Fernández, 1998). A continuación, la fórmula que la define (7).

$$\text{Sensibilidad} = \frac{TP}{TP + FP} \tag{7}$$

Especificidad:

La métrica especificidad es "la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de falsos positivos (FPR)", según (López de Ullibarri & Píta Fernández, 1998). A continuación, la fórmula que la define (8).

$$\text{Especificidad} = \frac{TN}{TN + FP} \tag{8}$$

Precisión (Accuracy)

Precisión o *accuracy*, es la relación entre el número de predicciones correctas y el número total de muestras de entrada, según (Mishra, 2018), como se indica en (9).

$$\text{Precisión} = \frac{Número\ de\ predicciones\ correctas}{Total\ de\ número\ de\ predicciones\ realizadas} \tag{9}$$

Funciona bien, solo si hay igual número de muestras pertenecientes a cada clase. Por ejemplo, considere que hay 98% de muestras de la clase A y 2% de la clase B, en el set de entrenamiento. Entonces, el modelo puede obtener fácilmente una precisión de entrenamiento del 98% con solo predecir cada muestra de entrenamiento perteneciente a la clase A, según (Mishra, 2018).

Cuando se ensaya el mismo modelo en un equipo de prueba con 60% de muestras de la clase A y 40% de muestras de la clase B, la precisión de la prueba se reduce al 60%. La precisión de la clasificación es grande, pero puede dar una falsa sensación de lograr una alta precisión, seegún lo indicado en (Mishra, 2018).

El verdadero problema surge cuando el costo de la clasificación errónea de las muestras de las clases menores es muy alto. Si se trata con una enfermedad rara pero fatal, el costo de no diagnosticar la enfermedad de una persona enferma es mucho más alto que el costo de enviar a una persona sana más pruebas, según lo indicado en (Mishra, 2018).

Curva ROC:

Las curvas ROC se desarrollaron en los años cincuenta como herramientas para el estudio de detección e interpretación de señales de radar (Burgueño et al., 1995). Los autores definen que la curva ROC es un gráfico en el que se observan todos los pares sensibilidad/especificidades, resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. En el eje de coordenadas "*y"* se sitúa la sensibilidad o fracción de verdaderos positivos.

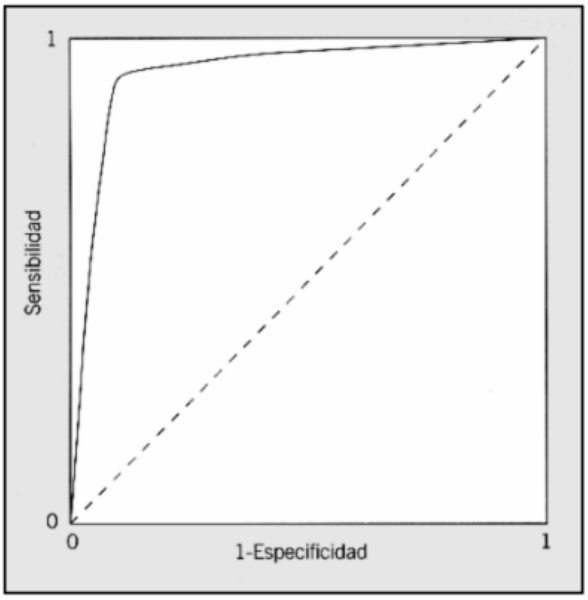

Figura No 1. Curva ROC. Tomada de (Burgueño et al., 1995).

Las curvas ROC son índices de la exactitud diagnóstica y proporcionan un criterio unificador en el proceso de evaluación de una prueba, debido a sus diversas aplicaciones (Burgueño et al., 1995). Los autores definen las siguientes ventajas del uso de curva ROC:

- Representación fácil y comprensible de la capacidad de discriminación de la prueba en todo el rango de puntos de corte.
- Gráficas simples y fáciles de interpretar visualmente.
- Como está incluido todo el espectro de puntos de corte, no requieren un nivel de decisión particular.
- Son independientes de la prevalencia, ya que la sensibilidad y la especificidad se obtienen en distintos subgrupos. Por tanto, no es necesario tener cuidado para obtener muestras con prevalencia representativa de la población. De hecho, es preferible generalmente tener igual número de individuos en ambos subgrupos.
- Proporcionan una comparación visual directa entre pruebas en una escala común, mientras que otro tipo de gráficos, como los diagramas de puntos o los histogramas de frecuencias, requieren diferentes gráficos cuando difieren las escalas.
- La especificidad y la sensibilidad son accesibles en el gráfico, en contraste con los diagramas de puntos y los histogramas.

Complementariamente, el parámetro que permite evaluar la prueba basada en los resultados es el área bajo la curva (*Area Under Curve* - AUC). Esta área puede interpretarse como la probabilidad de que la prueba clasifique las actividades correctamente.

Media Armónica:

Según (Universidad Pedagógica y Tecnológica de Colombia, 2004), la media armónica se define como el recíproco de la media aritmética de los recíprocos, como se indica en (10).

$$\mathrm{MA} = \frac{1}{\frac{1}{n}\left(\frac{1}{X_1}+\frac{1}{X_2}+\cdots+\frac{1}{X_n}\right)} \qquad (10)$$

El empleo de la media geométrica o de la armónica, equivale a una transformación de la variable en log X o 1/X, según (Universidad Pedagógica y Tecnológica de Colombia, 2004).

3. Trabajos relacionados

Producto del análisis de la revisión de la literatura, se han identificado una serie de trabajos relacionados con el HAR, en cuanto a: 1) la implementación de técnicas de preprocesamiento (Ronao & Cho, 2016), (Capela et al., 2015), (Gudivada et al., 2017), (Ren & Malik, 2003), (Galván-Tejada et al., 2016) y (Eddy, 1998); 2) la aplicación de técnicas basadas en aprendizaje supervisado y no supervisado (Shah, 2020), (Nettleton et al., 2010), (Caruana & Niculescu-Mizil, 2006) y (Mejia-Ricart et al., 2018); y 3) los distintos ámbitos de desarrollo del HAR que permitan el procesamiento de los datos extraidos de los sensores para resolver problemas relacionados con las actividades humanas en diferentes contextos (Crandall & Cook, 2013), (L. Chen et al., 2012), (Hoey et al., 2011) y (Fahad et al., 2015). Cada uno de estos referentes se abordará en detalle a continuación.

3.1 Implementación de técnicas de preprocesamiento en el har

Con el rápido desarrollo de la tecnología actual y la popularización de los teléfonos inteligentes, la detección ubicua se ha convertido en un campo de investigación y su propósito universal es extraer conocimiento a partir de los datos obtenidos por los sensores. La clave para un HAR exitoso es: 1) efectuar un apropiado proceso de representación de características de los datos recolectados mediante los sensores, 2) seleccionar las características mas representativas del conjunto de datos y 3) entrenar, evaluar e implementar clasificadores a partir de los subconjuntos de datos representados en cacarcterísticas. En (Ronao & Cho, 2016), se realiza un

estudio basado en la selección de características, allí se propone una red neuronal convolucional (*Convolutional Neural Networks - CNN*) que permite la extracción de características y su clasificación, usando sensores de teléfonos inteligentes. Una CNN multicapa, contiene membranas alternas de convolución y agrupación, las capas externas permiten la captación de las entradas, mientras que las capas internas tienen funciones de transformación. En (Ronao & Cho, 2016) se muestra como las diferentes arquitecturas de una CNN afectan el rendimiento general y como este sistema no requiere preprocesamiento avanzado o tediosas funciones de construcción manual, y puede superar a otros algoritmos de próxima generación en el campo HAR.

Por otra parte, en (Capela et al., 2015) determinaron las características de las señales que son más adecuadas para el reconocimiento de actividades humanas, utilizando teléfonos inteligentes, portados en la cintura por diferentes individuos. Capela realizó la selección de características independiente del clasificador. La identificación de subconjuntos de características que mejoraron la clasificación de actividades y a su vez mejoraron los modelos de monitoreo de movilidad para su uso en clasificadores futuros. Los subconjuntos de características con un rendimiento de clasificador similar al conjunto de características completo deberían reducir la carga computacional, facilitando así las implementaciones en tiempo real. Esta investigación es un paso importante en el desarrollo de un sistema HAR preciso y robusto para poblaciones diversas.

La calidad de los datos, según (Gudivada et al., 2017), juega un papel fundamental en el preprocesamiento. La adquisición y la verificación son los desafíos más importantes en las aplicaciones de uso intensivo de datos, por tanto, se debe tener en cuenta el grado de adecuación de los datos para un propósito determinado, es decir: que estén completos, sean coherentes, sin

duplicaciones, que sean precisos y oportunos. La aplicación de prácticas y controles relevantes, que mejoran la calidad de los informes, esto se conoce como calidad de los datos.

El análisis presentado por (Galván-Tejada et al., 2016), propone la recopilación de datos de diferentes fuentes utilizando un dispositivo y el uso de una versión modificada del algoritmo *Adaptive Boosting (AdaBoost)* para la selección de características. Un trabajo similar propuso (Eddy, 1998) en el uso de modelos ocultos de *Markov* (HMM) para clasificar ciertas actividades. Un ejemplo relevante de esta técnica es el propuesto por Sean Eddy, quien construyó un modelo denominado, *Discriminative Conditional Restricted Boltzmann Machine* (DCRBM). Este modelo combina un enfoque discriminatorio con las capacidades del *Conditional Restricted Boltzmann Machine* (CRBM). El modelo permite el descubrimiento de componentes procesables de Predicados esenciales de interacción social (ESIP) para entrenar el modelo DCRBM y usarlo para generar datos de bajo nivel, correspondientes al ESIP con un alto grado de precisión.

El proceso de extracción de características se emplea para representar un patrón, utilizando varios valores derivados o particularidades que pretenden ser informativas y no redundantes. Este proceso podría fácilmente resultar en conjuntos muy grandes de características que describen un patrón. Sin embargo, no todas las características extraídas serán útiles para discernir entre diferentes tipos de patrones, por lo que también es necesaria la selección de peculiaridades para determinar qué conjunto de características podría clasificar con precisión diferentes actividades. En (Galván-Tejada et al., 2016) se implementó un análisis de selección de particularidades que incluyó el uso de un algoritmo genético, la selección hacia adelante y los pasos de eliminación hacia atrás, y un algoritmo de bosque al azar (RF) o red neuronal (NN).

Otra importante subcategoría que se ha identificado en la implementación de técnicas de preprocesamiento en el HAR es la segmentación. En (Ren & Malik, 2003) se plantea un modelo

de clasificación para la segmentación a partir de señales Gestalt, en dicho estudio se toma una base de datos de imágenes humanas, posteriormente realiza una segmentación como ejemplo positivo. Para ejemplos negativos, emparejó aleatoriamente la imagen de un humano segmentada a una imagen diferente. Para la realización del clasificador, Ren utilizó *simple logistic regresión classifier*, con este clasificador obtuvo los mejores resultados. Para extraer más información sobre esta combinación de funciones, examinaron las distribuciones, tanto en los ejemplos positivos y los ejemplos negativos, verificando si son linealmente separables. Encontraron que las características normalizadas son aproximadamente gaussianas distribuidas y un clasificador lineal se ajusta bien a los datos. Una forma de evaluar el modelo fue observar los resultados finales de la segmentación con las métricas de precisión y *recall*.

3.2 Técnicas de basadas en aprendizaje supervisado y no supervisado aplicadas para el har

El aprendizaje supervisado es la metodología que más importancia tiene en el procesamiento de datos. Shah (Shah, 2020) en su libro *Supervised Learning*, en el capítulo dos (2) se muestran una serie de técnicas basadas en el aprendizaje supervisado, como las máquinas de soporte vectorial, las cuales según su estudio, tienen un buen desempeño particularmente en el procesamiento de datos multimedia, también considera el clasificador del vecino más cercano como una de las técnicas más populares en multimedia porque hace énfasis en la similitud. Nettleton (Nettleton et al., 2010) aplica técnicas de aprendizaje supervisado para extraer información útil, ya que frecuentemente los datos de prueba contienen ruido. Es decir, la calidad de estos puede verse disminuida por errores o desviaciones producidos en la fase de recopilación de la información, como consecuencia del error humano en la traducción de información o debido a las limitaciones en la tolerancia del equipo de medición. Esto puede

evidenciar errores en los valores de los atributos (ruido de atributo) o en la clase de las instancias. Generalmente, los datos ruidosos pueden sesgar el proceso de aprendizaje, lo que dificulta que los algoritmos de aprendizaje formen modelos precisos a partir de los datos. Por lo tanto, el desarrollo de técnicas de aprendizaje que se ocupen de forma eficaz y eficiente con este tipo de información es un aspecto clave en el aprendizaje automático. Nettleton realizó dos grupos de técnicas, el grupo uno compuesto por *Naïve Bayes* (NB) y *decision tree*, representa técnicas más robustas al ruido, y el grupo dos compuestos por *IBk* y SMO (*support vector machine*), representan técnicas que se proponen serán más sensibles al ruido. Al finalizar el estudio se concluyó que el mejor clasificador para altos porcentajes de ruido en datos de entrenamiento y prueba es SMO, seguido de NB.

En (Caruana & Niculescu-Mizil, 2006) presentan los resultados a gran escala de una comparación empírica de 10 algoritmos de aprendizaje supervisado (*SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees* y *boosted*), usando ocho (8) criterios de desempeño (*accuracy, F-score, Lift, ROC Area, average precision, precision/Recall break-even point, squared error* y *cross-entropy*). Para este trabajo comparativo, Caruana concluyó que los métodos de aprendizaje como *boosting, random forests, bagging y SVMs* logran un rendimiento excelente a la hora medir la precisión de los algoritmos de aprendizaje supervisado. De los algoritmos de aprendizaje más antiguos, *feedforward neural nets* tiene mayor rendimiento y es más competitivo que algunos de los métodos más nuevos, especialmente si los modelos no son calibrados después del entrenamiento.

En el estudio de (Mejia-Ricart et al., 2018) se propone un enfoque sin supervisión, con el fin de revelar patrones en los datos que podrían usarse como actividades primitivas para un mayor reconocimiento de actividades de nivel superior. Mejia recolectó los datos del sensor

usando un teléfono inteligente en un bolsillo lateral y el reloj inteligente sujeto a la muñeca izquierda. Ambos dispositivos fueron configurados para recolectar lecturas por medio del acelerómetro, giroscopio y podómetro. La frecuencia de muestreo para las lecturas del acelerómetro fue de 20 Hz, lo que equivale a una muestra cada 50 ms. Para el giroscopio, las lecturas se recopilaron cada 100 ms (10 Hz) y la frecuencia de muestreo de recopilación de datos de sensores se hizo mediante ventanas de tiempo deslizantes de dos (2) seg. Para la clasificación se usaron 6 algoritmos diferentes de *clustering*. Como consecuencia, en comparación con el conjunto de datos etiquetados, los resultados obtenidos de la agrupación no coincidieron con ninguna de las acciones primitivas, el *Average Linking* y *Mean Shif*t no lograron crear clústeres lo suficientemente similares para ser considerados.

3.3 Ámbito de desarrollo en har

La incorporación de los procesos de toma de decisión de manera proactiva e inteligente a las necesidades específicas del hogar es lo que produce un SH. Debido a esto, las comunidades de investigación, médica y empresarial han sido muy creativas para aprovechar este concepto para sus diversas necesidades. En (Crandall & Cook, 2013), se plantea que el área con la mayor factibilidad a largo plazo, para la comercialización de SH, es el cuidado de la salud. Puesto que, para la comunidad, el cuidado de la salud y la capacidad de supervisar de forma remota a los adultos mayores en su casa es de gran importancia y mitiga las dificultades que se le presenta a esta comunidad, debido a situaciones propias del envejecimiento. Las SH se ajustan a los residentes, sin afectar de manera negativa su estilo de vida. El sistema debe tomar información sobre el entorno del hogar e intentar construir modelos sobre las actividades e intereses de sus habitantes. Para hacer viviendas inteligentes capaces de apoyar este objetivo, la comunidad investigadora se ha centrado en desarrollar tecnologías para la detección de las actividades de la

vida diaria (ADL) como el seguimiento a los residentes (Crandall, 2011), la identificación de los mismos (Crandall & Cook, 2010), el desarrollo del historial médico (Cook, 2006), la interacción social (Cook et al., 2010) y la evaluación mental de los moradores de dichos hogares (Seelye et al., 2013), entre muchos otros aspectos.

El análisis a largo plazo de las actividades puede proporcionar información con respecto a la realización de tareas básicas, como: comer, dormir y cocinar, entre otras. El monitoreo de las personas y su entorno se recopila mediante sensores implementados en múltiples ubicaciones en una casa inteligente. Desde el enfoque basado en el reconocimiento de actividades, se aplican modelos ontológicos y de razonamiento semántico, que según Chen (L. Chen et al., 2012) se pueden combinar. Otro enfoque basado en el conocimiento, es el de la información de interacciones del usuario, que dentro del contexto se combina con los sensores disponibles en el hogar inteligente, para generar un modelo de decisión de Markov que permita el reconocimiento de las actividades de la vida diaria y su incidencia en el bienestar de los adultos mayores (Hoey et al., 2011).

En (Fahad et al., 2015) se describe el *dataset* Kyoto con las instancias y la información de las actividades, además se muestra un balance de un modelo propuesto y la comparación con otros conjuntos de datos, que se pueden observar en la siguiente tabla No. 4 donde también se puede apreciar que el *dataset* de kyoto tiene los mejores resultados con relación al modelo propuesto por Fahad.

Tabla No. 4

Comparación del enfoque propuesto en (Fahad et al., 2015) *con ET-KNN, sin selección de caracteristicas y balanceo de datos, utilizando los dataset kyoto y kasteren*

Datasets	Folds	Approach	Precision (%)	Recall (%)	F1score [0, 1]	Accuracy (%)
Kyoto1	Three folds	Proposed approach	**97.33**	**96.67**	**0.97**	**96.67**
		ET-KNN	95.85	95.00	0.95	95.00
	One day out	Proposed approach	**98.11**	**97.44**	**0.97**	**97.44**
		ET-KNN	97.22	96.15	0.95	96.15
Kyoto7	Three folds	Proposed approach	**79.00**	**76.79**	**0.77**	**80.00**
		ET-KNN	72.35	69.59	0.69	74.15
	One day out	Proposed approach	**76.60**	**80.09**	**0.77**	**81.00**
		ET-KNN	70.72	75.74	0.71	76.07
Kasteren7	Three folds	Proposed approach	**90.10**	**93.11**	**0.91**	**94.21**
		ET-KNN	89.89	87.77	0.87	92.27
	One day out	Proposed approach	**94.13**	**94.13**	**0.94**	**95.28**
		ET-KNN	93.41	89.66	0.90	93.06
Kasteren10	Three folds	Proposed approach	88.80	88.40	0.87	92.54
		ET-KNN	**90.70**	**84.31**	**0.85**	**90.04**
	One day out	Proposed approach	**88.10**	**89.14**	**0.88**	**92.00**
		ET-KNN	83.62	83.32	0.82	90.77

Fuente: (Fahad et al., 2015)

4. Metodología

En este capítulo se presentan diferentes metodologías relacionadas con el proceso de Minería de Datos, en aras de sustentar la metodología aquí utilizada. Con base en ella se muestra un diagrama del modelo funcional predictivo propuesto con su respectiva explicación y por último se enumeran los diferentes escenarios de experimentación que se han recreado para el desarrollo de la misma.

4.1 Metodologias propias de procesos basados en minería de datos

Para la implementación de procesos de minería de datos se han desarrollado diferentes metodologías, las tres más empleadas son: 1) el Descubrimiento de conocimiento en bases de datos - *Knowledge Discovery in Databases* – KDD (Fayyad et al., 1996); 2) el muestreo, exploración, modificación, modelado y evaluación - *Sample, Explore, Modify, Model and Assess* – SEMMA (Azevedo & Santos, 2005) y 3) el proceso estándar de la industria cruzada para la minería de datos - *Cross Industry Standard Process for Data Mining* - CRIPS-DM, definida en (Catley et al., 2009) y (Spruit et al., 2014).

4.1.1. Metodología KDD

La metodología KDD, definida en (Fayyad et al., 1996), permite el uso de diferentes métodos basados en minería de datos. En una primera etapa, se requiere la extracción de lo que se considera conocimiento, de acuerdo con especificaciones de medidas y umbrales, usando un conjunto de datos. Posterior a la extracción, en esta metodología, se ejecutan diferentes etapas relacionadas con el preprocesamiento, submuestreo y transformación de los datos. Esta metodología posee cinco cinco etapas, que se presentan en la figura No.2, las cuales se describen a continuación: La primera etapa, de selección, consiste en crear el conjunto de datos

o muestras de datos, sobre las cuales se aplicará la metodología; la segunda etapa; también denominada preprocesamiento, que consiste en la limpieza de los datos de destino para obtener información consistente; la tercera etapa consiste en la transformación de los datos, utilizando métodos de reducción de dimensionalidad; la cuarta etapa busca implementar mediante la minería de datos, patrones de interés en una forma de representación particular, dependiendo del objetivo (generalmente, predicción); y la quinta etapa, interpretación y evaluación, busca identificar patrones extraídos a partir de los datos y por medio de estos incorporar el conocimiento. También puede estar representado en informes o documentación a las partes interesadas.

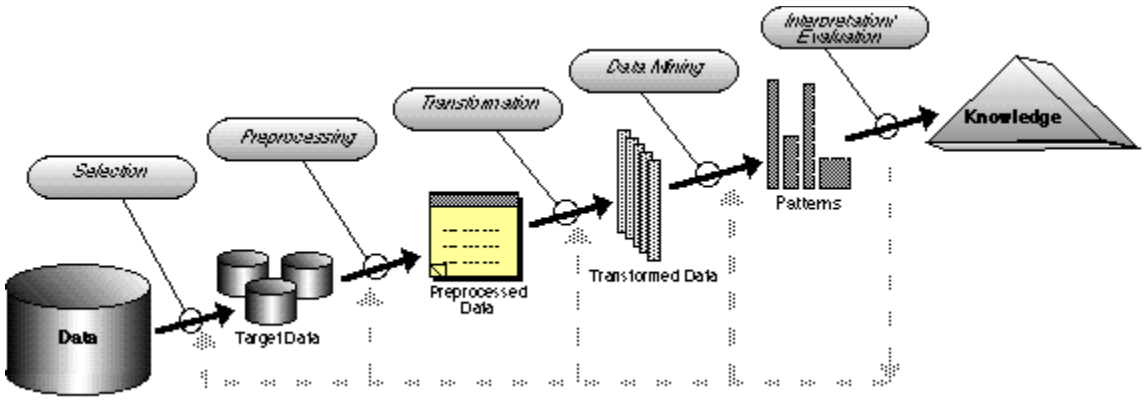

Figura No 2. Etapas de la metodología KDD
Fuente: (Fayyad et al., 1996)

La metodología KDD debe estar precedida por conocimiento previo, relevante y tener claridad sobre los requerimientos del usuario final.

4.1.2. Metodología SEMMA.

La metodología SEMMA, definida en (Azevedo & Santos, 2005), se emplea para el desarrollo de proyectos basados en minería de datos. SEMMA se presenta como un ciclo caracterizado en cinco (5) etapas: 1) la etapa de muestreo que permite la extracción de una suficiente y significativa cantidad de información, a partir de un gran conjunto de datos, con el propósito de que sean fácilmente manejables; 2) consiste en la exploración de los datos mediante la búsqueda de tendencias imprevistas y anomalías; 3) esta etapa consiste en la modificación de los datos mediante la creación, selección y transformación de las variables para enfocar el proceso de selección del modelo, 4) consiste en modelar la información permitiendo que el software busque automáticamente una combinación de datos que predijeren de manera confiable un resultado deseado. 5) esta etapa evalúa los datos estimando la utilidad y confiabilidad de los hallazgos del proceso de minería de datos y qué tan bien funciona. SEMMA ofrece un proceso de fácil comprensión, permitiendo un desarrollo organizado y el mantenimiento de proyectos de minería de datos.

4.1.3. Metodología CRIPS-DM.

La metodología CRIPS-DM, definida en (Catley et al., 2009) y (Spruit et al., 2014), es usada para modelar sistemas basados en análisis inteligentes de la información, en contextos de minería de datos, organizados por ventanas de tiempo. Esta metodología consta de seis (6) fases las cuales se describen de manera sucinta a continuación: 1) Comprensión Empresarial: esta fase determina los objetivos del negocio, además de la situación a evaluar al incluir la minería de datos se desarrolla el plan de desarrollo del proyecto, 2) Comprensión de datos: esta fase hace referencia a la agrupación de los tipos de datos dentro del *dataset* y al entendimiento de estos,

teniendo en cuenta la calidad de los datos. 3) Preparación de datos: en esta etapa se deben seleccionar las variables a usar, además de, agrupar los datos necesarios para la herramienta de modelado, incluye la limpieza, integración y formato del *dataset*, 4) Modelado: en esta etapa se elige la técnica de modelado de minería de datos a usar con el fin de obtener resultados óptimos, se genera el diseño de pruebas y finalmente el modelo a evaluar; 5) Evaluación: este proceso evalúa si uno o más modelos cumple con los objetivos de la etapa inicial. 6) Implementación: esta etapa usa los modelos resultantes del proceso de minería de datos y se presentan en forma de descripciones que sean de fácil entendimiento (Nuraeni et al., 2019).

Figura No 3. Metodología CRISP-DM**.***Fuente:* (Nuraeni et al., 2019)

4.2 Metodología empleada

Al hacer una comparación de las metodologías KDD y SEMMA, se puede afirmar que son equivalentes con respecto a las etapas de construcción del modelo, identificando un proceso de selección, luego el procesamiento y transformación de los datos, también se evidencia la etapa del modelo en la que se aplican diferentes técnicas de minería de datos, finalmente se evalúa de

acuerdo con las métricas de calidad establecidas. Se puede exponer que las etapas de la metodología SEMMA son una puesta en marcha práctica de KDD.

Comparar la metodología KDD con CRISP-DM no es tan simple como en el SEMMA. No obstante, se puede intuir que CRISP-DM incorpora ciclos que hacen referencia a la metodología KDD como: la fase de comprensión empresarial es proporcional a comprensión del conocimiento previo relevante y los objetivos del usuario final, la fase de comprensión de datos en la metodología CRISP-DM es equivalente a la combinación de las etapas de selección y preprocesamiento en la metodología KDD. La preparación de datos puede ser proporcional con la etapa de transformación, el modelado con la de minería de datos y finalmente la etapa de evaluación se identifica con la interpretación/evaluación. En la tabla No. 5 se evidencia un comparativo entre las metodologías mencionadas.

Tabla No 5.

Comparativo de las etapas de las metodologías KDD, SEMMA Y CRISP-DM

KDD	SEMMA	CRISP-DM
Conocimiento Previo	N/A	Entendimiento Empresarial
Selección	Muestra	Comprensión de datos
Preprocesamiento	Exploración	
Transformación	Modificación	Preparación de datos
Minería de datos	Modelo	Modelamiento
Interpretación/Evaluación	Evaluación	Evaluación
N/A	N/A	Implementación

En conclusión SEMMA y CRISP-DM pueden verse como una implementación del proceso KDD, de acuerdo a lo descrito por (Fayyad et al., 1996). Se puede entrever que CRISP-DM es más completo que SEMMA. Sin embargo, al analizarlo se puede apreciar la integración de la comprensión de los datos, en las etapas de muestra y exploración de SEMMA, debido a que los datos no se pueden manipular a menos que exista una comprensión de todos los aspectos.

La metodología usada en esta investigación para la predicción de ADL, es KDD, dado que permite la extracción y análisis de información, previamente desconocida y potencialmente útil. Esta metodología consiste en extraer patrones que permiten al usuario obtener información a partir de conjuntos de datos masivos. Esto mediante el preprocesamiento y la minería de datos, seguidamente se presentan los resultados mediante un modelo entrenado, el cual estimará la actividad que realiza una persona en un ambiente *indoor*. Finalmente, en el proceso de evaluación se aplican métricas de calidad que buscan generar un modelo predictivo con los mejores resultados obtenidos, teniendo así un modelo validado.

4.3 Modelo funcional predictivo

El proceso de construcción del modelo predictivo para el reconocimiento de ADL a partir del *dataset* CASAS Kyoto, implicó el planteamiento de un proceso experimental, dividido en una serie de fases (ver figura No. 4), es decir: 1) integración y depuración, 2) agrupamiento de instancias, 3) aplicación de técnicas de representación de características por subconjunto de datos, 4) entrenamiento y prueba de modelos para la clasificación, y (5) evaluación de las métricas de calidad de los modelos para identificar con cual hibridación de técnica se generaron los mejores resultados, en cuanto a tasa de aciertos. A continuación, se detallan cada una de las fases antes mencionadas.

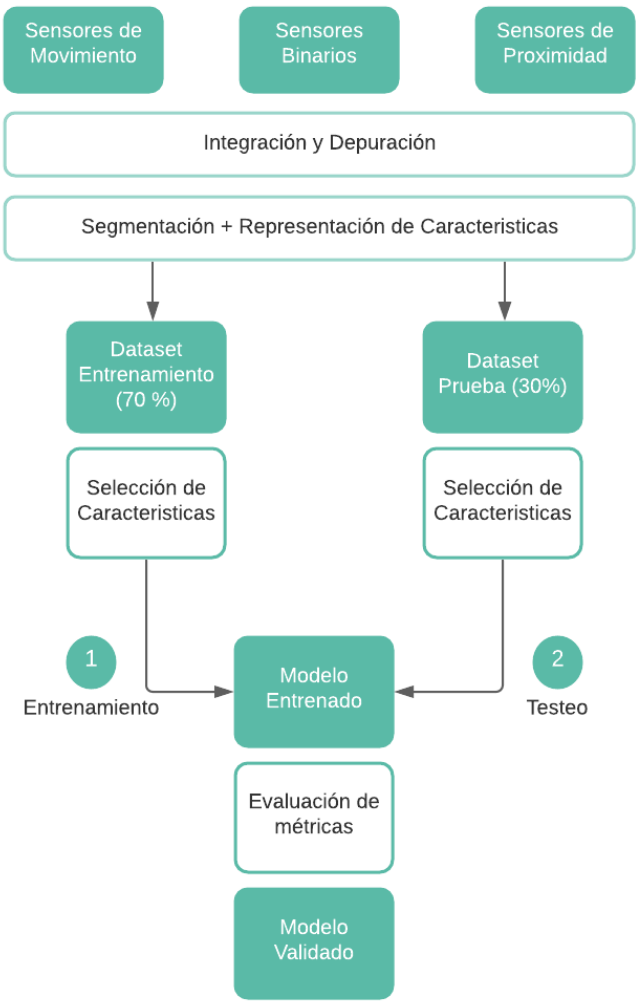

Figura N.4. Proceso de construcción del modelo funcional predictivo

4.3.1. Integración y depuración.

Para generar *dataset* de ADL *Activities*, los investigadores del laboratorio CASAS reclutaron a 20 voluntarios participantes para realizar cinco (5) actividades las cuales son: efectuar llamadas telefónicas, lavarse las manos, cocinar, comer y limpiar. La información fue recolectada a través de los siguientes sensores: Sensores de movimiento, Sensores para la utilización de elementos de cocina, Sensor en el contenedor de las medicinas, sensor en utensilios de cocina, directorio telefónico, sensor de gabinetes, sensor de agua, sensor de

encendido de la cocina y uso del teléfono. Para la integración de los datos, primero se generó un *dataset* constituido por 120 archivos que corresponden a mediciones de 86 días de las interacciones de los individuos *vs* sensores en el ambiente, esas instancias fueron ordenadas consecutivamente por la columna *TimeStamp,* esta integración se realizó debido a que cada uno de los archivos tienen la misma estructura. La estructura de cada uno de los archivos está constituida por: fecha, *TimeStamp,* sensor, estado, inicio y finalización de la actividad. Producto de esto, se obtuvo un primer *dataset row* con un total de 6.425 instancias de datos.

Para la construcción del *dataset* preprocesado, se tomó el *dataset row* en el cual, teniendo en cuenta el *TimeStamp* se generó una representación en columnas de los estados de cada uno de los sensores de la respectiva línea de tiempo, indicando de igual forma, la actividad que se estaba ejecutando en dicha línea. El *dataset* preprocesado está constituido por cinco (5) tipos de características o columnas de datos (las cuales están representadas en la tabla No. 6) y recopilan un total de 26 características.

Tabla No. 6

Descripción de las características del dataset Preprocesado

#	Característica	Cant	Descripción
1	TimeStamp	1	Contiene la fecha y hora de la secuencia de actividades.
2	Sensores de movimiento	11	Identificados así: M01, del M07 al M09, del M13 al M18 y el M23. Cada uno toma valores ON y OFF.
3	Sensores binarios	10	Identificados así: del I01 al I08 y D01. Cada uno toma valores ABSENT y PRESENT.
4	Sensores de proximidad	3	Identificados así: AD1-A, AD2-B y AD1-C. Cada uno toma valores numéricos.
5	Uso telefónico	1	Toma valores de START y END.
	Total	**26**	

4.3.2. Agrupamiento de instancias y aplicación de técnicas de representación de características por subconjunto de datos

Luego del proceso descrito anteriormente, en el que se obtiene el *dataset* preprocesado, se continua con el agrupamiento de las instancias de datos; para ello se procedió a segmentar el *dataset* con ventanas deslizante de tres (3) segundos, mientras que, para el proceso de generación de nuevas características se tomaron los valores de los atributos que corresponden a los sensores de proximidad (AD1-A, AD2-B y AD1-C), aplicando funciones de agregación tales como: desviación estándar, curtosis, media, máximo, mínimo, sesgo y rango. Producto de esto se obtuvieron 21 nuevas características, generando un *dataset* segmentado con 45 características incluyendo la actividad y 5.736 instancias. Ver tabla No 7.

Tabla No 7.

Descripción de las características del dataset segmentado

#	Característica	Cant	Descripción
1	TimeStamp	1	Contiene la fecha y hora de la secuencia de actividades.
2	Sensores de movimiento	11	Identificados así: M01, del M07 al M09, del M13 al M18 y el M23. Cada uno toma valores ON y OFF.
3	Sensores binarios	10	Identificados así: del I01 al I08 y D01. Cada uno toma valores ABSENT y PRESENT.
4	Sensores de proximidad	21	Por cada una de las TRES características iniciales de proximidad (AD1-A, AD2-B y AD1-C) se aplicaron las funciones de agregación: desviación estándar, media, rango, max, min, curtosis y sesgo.
5	Uso telefónico	1	Toma valores de START y END.
6	Actividad	1	Limpiar, Cocinar, Comer, Lavarse las manos y Llamada telefónica
	Total	**45**	

El conjunto de datos después del proceso de segmentación está desbalanceado, es decir, el número de instancias de cada una de las actividades están en diferentes proporciones. Una

solución a esto es el balanceo de las clases. Para ilustrar el funcionamiento de esta técnica se deben identificar las clases minoritarias para luego sobre muestrear, es decir, tomar una muestra del conjunto de datos y considerar sus vecinos más cercanos, para crear un punto de datos más sintético. Para el balanceo de las clases de utilizó la técnica SMOTE (*Synthetic Minority Oversampling Technique*) (Chawla et al., 2002) obteniendo como resultado un *dataset* balanceado como lo muestra la Figura 5.

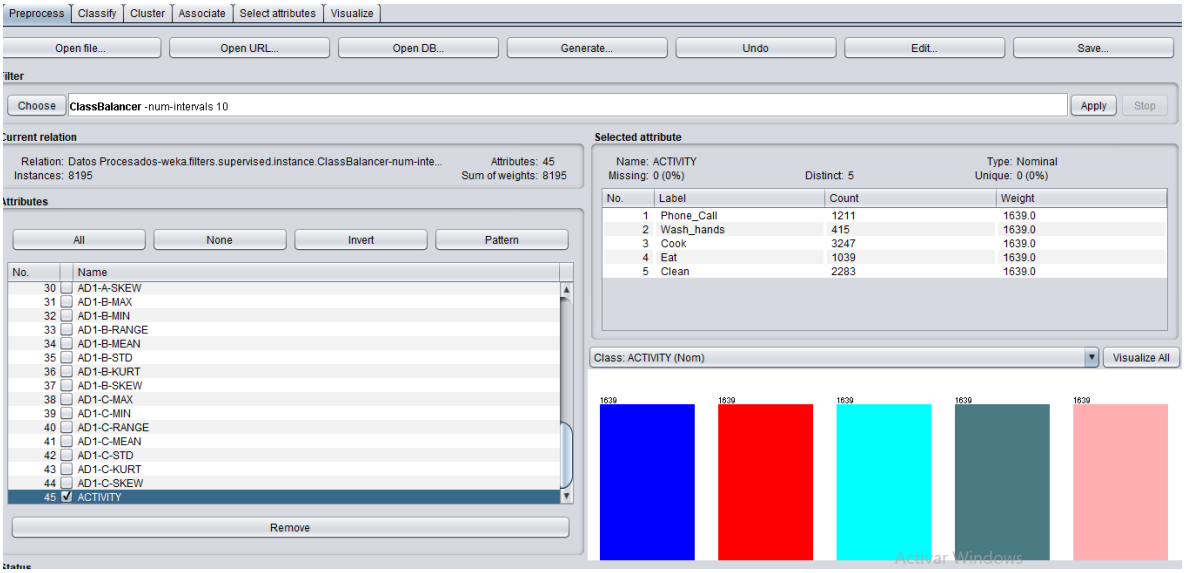

Figura No. 5 Clases del Dataset Balanceado. Fuente: Elaboración propia

4.3.3. Entrenamiento y prueba de modelos para la clasificación

Al realizar el balanceo de las clases como se evidencia en la Figura No. 6, se puede observar la simetría en cada una de las instancias.

Seguidamente se generaron dos subconjuntos de datos el primero para el entrenamiento del modelo y segundo para pruebas, con una proporción 70-30, dado que la literatura (López Saca et al., 2018) indica 70% de los datos para el entrenamiento y 30% para *test*, con registros diferentes en cada *dataset* .

Para el proceso de aplicación de técnicas de selección de características se identifica el nivel de incidencia que tienen los atributos del *dataset* con respecto al proceso de identificación del criterio de clase. Este proceso se llevó a cabo con la aplicación de diferentes técnicas de selección de características tales como: *Gain Ratio, Chi square, Info Gain, OneR, ReliefF y Symmetrical Uncert*. Este proceso busca una reducción en el número de las características con el fin de disminuir el tiempo computacional requerido en la construcción del modelo definitivo.

Posteriormente, se realiza el entrenamiento del modelo usando los atributos con mejores resultados de acuerdo con la selección de características. con el propósito de que este pueda predecir la actividad que realiza una persona de acuerdo con el conjunto de datos ingresados, en esta etapa se obtienen varios modelos, los cuales posteriormente se les aplicarán diferentes métricas de calidad con la finalidad de encontrar el que mejor resultado arroje.

4.3.4. Evaluación de las métricas de calidad de los modelos

Finalmente, se realiza la evaluación de las métricas de calidad del modelo obtenido para identificar qué hibridación de técnica de clasificación con técnica de selección de características, arroja los mejores resultados en cuanto a la tasa de aciertos y luego se realiza la selección del modelo con los mejores resultados. Las técnicas de calidad que fueron evaluadas son las siguientes: *FPR,* Precisión, *Recall y ROC* Area.

Comprendida la metodología usada en este estudio, se procedió a recrear una serie de escenarios de experimentación, usando diferentes tipologías de técnicas de segmentación, selección de características y de clasificación. Cada uno de estos escenarios se explica detalladamente en la sección 5.

5. Experimentación

En este proyecto se han recreado los siguientes escenarios de experimentación: 1) la evaluación del modelo con el *dataset* preprocesado; 2) la evaluación del modelo utilizando el *dataset* segmentado; 3) la evaluación del modelo utilizando el *dataset* segmentado y balanceado, y 4) la evaluación del modelo hibridando técnicas de selección de características y clasificación con el *dataset* segmentado y balanceado.

5.1 Escenario no 1: evaluación del modelo utilizando el dataset preprocesado

En este escenario, el *dataset* procesado (constituido por 45 características incluido el criterio de clase, para mayores precisiones ver la tabla 7 de la sección 4.3.2) fue evaluado usando las siguientes diecinueve (19) técnicas de clasificación: *Classification Via Regression* (Ruan et al., 2014), *Random SubSpace* (Ho, 1998), *Bagging* (Breiman, 1996), *Random Forest* (Nagalla et al., 2017), *Attribute Selected, J48* (Salzberg, 1994), *OneR* (Kumar et al., 2013), *LMT* (W. Chen et al., 2017), *REP Tree* (Chandra & Maheshkar, 2017), *Randomizable Filtered* (Asaju et al., 2017), *Random Tree* (Kalmegh, 2015), *JRip* (Rajput et al., 2011), *Iterative Classifier Optimizer, LogiBoost* (Cai et al., 2006), *Multi Class* (Qian et al., 2010), *Multilayer Perceptron* (Suykens & Vandewalle, 1999), *Simple Logistic* (Khalajzadeh et al., 2014) *e IBK* (Choudhury & Bhowal, 2015), Ver tabla 8.

Tabla No. 8
Resultados de la evaluación de técnicas de clasificación con el dataset completo

#	Clasificador	FPR	Precision	Recall	ROC Area
1	**Classification Via Regression**	**1,20%**	**97,60%**	**97,60%**	**99,70%**
2	Random SubSpace	4,10%	89,80%	**89,90%**	98,60%
3	Bagging	5,20%	85,10%	**86,40%**	97,40%

4	RandomForest	4,10%	86,90%	**87,40%**	97,20%
5	Attribute Selected	2,70%	91,60%	**91,80%**	96,90%
6	J48	2,70%	91,20%	91,60%	96,80%
7	OneR	2,80%	95,10%	94,80%	96,00%
8	LMT	5,50%	84,20%	85,20%	96,00%
9	REP Tree	5,70%	83,00%	84,50%	95,70%
10	RandomCommittee	4,20%	86,90%	87,20%	91,50%
11	Random Tree	4,10%	86,60%	87,00%	91,40%
12	JRip	9,00%	84,40%	83,90%	91,30%
13	Iterative Classifier Optimizer	23,80%	61,20%	56,90%	78,30%
14	LogitBoost	23,80%	61,20%	56,90%	78,30%
15	Logistic	26,20%	56,20%	53,50%	76,50%
16	Multi Class	25,90%	56,00%	53,50%	74,80%
17	Multilayer Perceptron	25,00%	53,50%	52,70%	74,50%
18	Simple Logistic	26,30%	58,00%	53,40%	74,30%
19	IBK	25,70%	57,20%	52,90%	71,60%

Fuente propia del autor

Para determinar cuál de estas técnicas de clasificación generó mejores resultados, se evaluaron las diferentes métricas de calidad documentadas previamente en la sección 2.3.3. Se identificaron cinco (5) clasificadores con las mejores métricas en cuanto a ROC Area y *Recall*, todos con un ROC área superior o igual a 96,9%. Dichos clasificadores son: *Classification Via Regression*, *Random SubSpace*, *Bagging*, *RandomForest* y *Attribute Selected.*

Los mejores resultados se obtuvieron con *Classification Via Regression* cuyo *Recall* fue de 97,6 % indicando así que la tasa de detección de las diferentes actividades fue alta, con un ROC Area de 99,7 % lo que denota que la calidad de la evaluación fue apropiada y un FPR de 1,2 % muestra que la metrica tuvo una baja tasa de falsos positivos.

5.2 Escenario no 2: evaluación del modelo utilizando el dataset segmentado

En este escenario el *dataset* fue segmentado en ventanas de tiempo deslizantes de tres (3) segundos, pasando así de 6425 instancias del *dataset* preprocesado a 5736 instancias en el dataset segmentado. Se aplicaron las mismas técnicas de clasificación, anteriormente listadas en el primer escenario. Producto de ello, se identificaron los 10 clasificadores que arrojaron mejores resultados, ver tabla 9.

Tabla No. 9.
Resultados Técnicas de Clasificación dataset segmentado

#	Técnica Clasificación	FPR	Precision	Recall	ROC Area
1	***Classification Via Regression***	**0,10%**	**99,80%**	**99,80%**	**100,00%**
2	*OneR*	0,30%	99,50%	**99,50%**	99,60%
3	*Attribute Selected*	0,90%	97,10%	**97,10%**	98,50%
4	*J48*	0,90%	96,80%	**96,80%**	98,60%
5	*Random SubSpace*	3,80%	91,50%	91,10%	98,70%
6	*Random Forest*	3,20%	90,30%	90,40%	98,40%
7	*Bagging*	3,50%	89,80%	90,00%	98,80%
8	*RandomCommittee*	3,30%	89,80%	89,90%	95,90%
9	*REP Tree*	3,60%	89,00%	89,20%	97,80%
10	*LTM*	3,90%	88,30%	88,60%	97,10%

Para determinar cuál de estas técnicas de clasificación generó mejores resultados, se evaluaron diferentes métricas de calidad. Con el clasificador *Classification Via Regression*, el *Recall* más alto fue 99,80% y el FPR fue 0,10%. En la tabla 10 se evidencia la mejora de las métricas de calidad en los clasificadores con mejores resultados del escenario 1 y se realiza la comparación con los resultados obtenidos luego del proceso de segmentación.

Tabla No. 10
Comparación mejores resultados escenario 1 vs escenario 2

Clasificador	Escenario 1 *Dataset* preprocesado			Escenario 2 *Dataset* segmentado		
	FPR	**Recall**	**ROC Area**	**FPR**	**Recall**	**ROC Area**
Classification Via Regression	**1,20%**	**97,69%**	**99,70%**	**0,10%**	**99,80%**	**100,00%**
Random SubSpace	**4,10%**	**89,90%**	98,60%	**3,80%**	**91,10%**	**98,70%**
Bagging	5,20%	86,40%	97,40%	3,50%	90%	98,80%
RandomForest	4,10%	87,40%	97,20%	3,20%	90,40%	98,40%
Attribute Selected	2,70%	91,80%	96,90%	0,90%	97,10%	98,50%

Con esta comparación podemos denotar que hubo una mejora en todos los clarificadores, el ROC Area con mejores resultados del primer escenario pasó de 99,7% a 100% en el segundo escenario y la tasa de FPR disminuyó sustancialmente en 1,1% del escenario 1 al escenario 2.

Finalmente se puede señalar que los clasificadores con los cuales se obtuvieron las mejores tasas de ROC área y *recall* fueron *Classification Via Regression, OneR y Attribute Selected*. Otros clasificadores que anteriormente estaban ranqueados en las diez (10) primeros puestos, cambiaron su posición como lo podemos evidenciar con el clasificador *OneR*, este en el primer escenario se encontraba en la séptima posición y con proceso de segmentación pasa al segundo lugar, lo que indica que el proceso de segmentación favorece a este clasificador, prueba de ello es el ROC área de 99,6% y *recall* de 99,5%.

5.3 Escenario no 3: evaluación del modelo utilizando el dataset segmentado y balanceado

El balanceo se realiza con respecto al criterio de clase, la cantidad de instancias de datos por clase cambia, gracias a la aplicación de la técnica de balanceo denominada SMOTE, ver tabla 11.

Tabla No.11

Comparación dataset segmentado vs dataset segmentado y balanceado

Clases	*Dataset* segmentado				*Dataset* Segmentado y balanceado			
	Train		Test		Train		Test	
Realizar llamada telefónica	822	14,33%	389	15,81%	1147	20%	491	19,97
Lavarse las manos	297	5,17%	118	4,80%	1147	20%	492	20,01
Cocinar	2299	40,08%	948	38,55%	1147	20%	492	20,01
Comer	719	12,53%	320	13,01%	1147	20%	492	20,01
Limpiar	1599	27,89%	684	27,83%	1148	20.01%	492	20,01
Total	**5736**	**100%**	**2459**	**100%**	**5736**	**100%**	**2459**	**100%**

Fuente propia del autor

Podemos observar que en el *dataset* desbalanceado hay un número mayor de instancias de la clase cocinar y limpiar, estas tienen un 67,97% del total de instancias *referenciadas en la tabla*. Producto del balanceo se evidencia el mismo número de instancias en el conjunto de datos.

En este escenario, el *dataset* segmentado (con ventanas de tiempo deslizantes de tamaño fijo de 3 segundos) y balanceado (constituido por 45 características, incluido el criterio de clase) se aplicaron las mismas técnicas de clasificación, anteriormente listadas en el primer escenario. Producto de ello, se identificaron los 10 clasificadores que arrojaron mejores resultados, ver tabla 12.

Tabla No. 12
Resultados Técnicas de Clasificación dataset completo

#	Técnica Clasificación	FPR	Precision	Recall	ROC Area
1	***Classification Via Regression***	**0.28%**	**99.55%**	**99.95%**	**99.96%**
2	*OneR*	0.14%	99.75%	**99.75%**	99.80%
3	*Random SubSpace*	2.15%	95.06%	**94.75%**	99.47%
4	*Bagging*	3.42%	90.14%	90.28%	98.80%
5	*J48*	0.77%	97.62%	97.64%	98.75%
6	*Attribute Selected*	0.69%	97.90%	97.92%	98.74%
7	*Random Forest*	3.10%	90.93%	91.09%	98.35%
8	*REP Tree*	3,52%	89.39%	89.54%	97,56%
9	*RandomCommittee*	3,18%	90,49%	90,64%	95,84%
10	*LTM*	3,35%	89,39%	89,67%	97,33%

Fuente propia del autor

Para determinar cuál de estas técnicas de clasificación generó mejores resultados, se evaluaron las diferentes métricas de calidad. Con el clasificador *Classification Via Regression*, el área ROC más alta fue 99,96% y el *Recall* fue de 99,95%, y el FPR fue 0,28%. De igual manera podemos evidenciar una mejora en los diferentes clasificadores con respecto a los escenarios anteriores como lo muestra la tabla 13.

Tabla No. 13.
Comparación mejores resultados escenario 1, escenario 2 y escenario 3

Clasificador	Escenario 1 *Dataset* preprocesado			Escenario 2 *Dataset* segmentado			Escenario 3 *Dataset* segmentado y balanceado		
	FPR	Recall	ROC Area	FPR	Recall	ROC Area	FPR	Recall	ROC Area
Classification Via Regression	**1,20%**	**97,60%**	**99,70%**	**0,10%**	**99,80%**	**100,00%**	**0.28%**	**99,50%**	**99.96%**
OneR	2,80%	94,80%	96,00%	**0,30%**	**99,50%**	**99,60%**	**0.14%**	**99,75%**	**99.80%**
Attribute Selected	2,70%	91,80%	96,90%	**0,90%**	**97,10%**	**98,50%**	0.69%	**97,92%**	97.92%
J48	2,70%	91,60%	96,80%	0,90%	96,80%	98,60%	0.77%	97,64%	98.75%
Random SubSpace	**4,10%**	**89,90%**	**98,60%**	3,80%	91,10%	98,70%	2.15%	94,75%	99.47%
Random Forest	4,10%	87,40%	97,20%	3,20%	90,40%	98,40%	3.10%	91,09%	91.09%
Bagging	**5,20%**	**86,40%**	**97,40%**	3,50%	90,00%	98,80%	3.42%	90,28%	98.80%

Fuente propia del autor

Se ratifica el hecho de que los clasificadores con los cuales se obtuvieron las mejores tasas de ROC área y *recall* fueron *Classification Via Regression y OneR*. El clasificador *Random SubSpace* en el primer escenario estuvo en el segundo lugar con un *ROC Area de* 98,60%, en el segundo escenario estuvo en el quinto lugar con un *ROC Area* de 98,7%, y luego del balanceo de cargas se ubicó en el tercer lugar con *ROC Area* de 99.47%. Otros clasificadores que anteriormente estaban ranqueados en los diez (10) primeros puestos, cambiaron su posición como lo podemos evidenciar con el clasificador *OneR*, el cual en el primer escenario se encontraba en la septima posición y con proceso de segmentación pasa al segundo lugar, lo que indica que este proceso favorece a este clasificador.

Un clasificador como *OneR* genera un árbol de decisión de un nivel capaz de inferir una clasificación típicamente simple, pero precisa con respecto a las reglas de un conjunto de instancias, para este escenario se evidencian mejores tasas de ROC área y *Recall* debido a la aplicación de la técnica de segmentación, en el primer escenario tuvo una tasa de ROC área de 96,80% y de *Recall* de 94,80%, posteriormente la segmentación la tasa fue de 99,80% y 99,75% respectivamente, con esto podemos evidenciar el aumento en el nivel de calidad en el proceso de predicción.

Producto del proceso de segmentación en ventanas de tiempo deslizantes de tres (3) segundos se ha observado un notable mejoramiento en los tres (3) clasificadores con las tasas más altas, entre los que podemos destacar *Classification Via Regression, OneR y Random SubSpace* generando un incremento en las métricas de ROC área, *Recall y Precision*, de igual manera, también se observa una disminución en la tasa de falsos positivos del 0,92% con respecto al primer clasificador. Cabe destacar que, *OneR* presenta una disminución en la tasa de

falsos positivos de 2,4%, con lo que podemos indicar que éste tiene mejor desempeño con respecto a *Classification Via Regression*.

5.4 Escenario no 4: evaluación del modelo hibridando técnicas de selección y clasificación con el dataset segmentado y balanceado

Como primer paso en este escenario se estableció una numeración para cada uno de los atributos, ver tabla. 14. Luego, aplicando las técnicas de selección de características se evaluó su nivel de incidencia, ver tabla 15

Tabla No 14

Atributos del dataset

Id	Atributo	Id	Atributo	Id	Atributo
0	DATE	15	I04	30	AD1-B-MAX
1	M01	16	I05	31	AD1-B-MIN
2	M07	17	I06	32	AD1-B-RANGE
3	M08	18	I07	33	AD1-B-MEAN
4	M09	19	I08	34	AD1-B-STD
5	M13	20	D01	35	AD1-B-KURT
6	M14	21	asterisk	36	AD1-B-SKEW
7	M15	22	E01	37	AD1-C-MAX
8	M16	23	AD1-A-MAX	38	AD1-C-MIN
9	M17	24	AD1-A-MIN	39	AD1-C-RANGE
10	M18	25	AD1-A-RANGE	40	AD1-C-MEAN
11	M23	26	AD1-A-MEAN	41	AD1-C-STD
12	I01	27	AD1-A-STD	42	AD1-C-KURT
13	I02	28	AD1-A-KURT	43	AD1-C-SKEW
14	I03	29	AD1-A-SKEW	44	Actividad

Fuente propia del autor

Tabla No 15

Prioridad de las características según técnica de selección de atributos

Id	GainRatio	Id	InfoGain	Id	OneR	Id	ReliefF	Id	SymmetricalUncert	Id	ChiSquared
21	0,3294	0	2,0407	0	99,546	23	0,0597	0	0,4758	0	22,943
5	0,3173	5	0,1328	5	45,571	26	0,0597	5	0,1080	5	12,793
3	0,3133	26	0,0955	33	44,334	24	0,0597	26	0,0793	24	63,906

Id	GainRatio	Id	InfoGain	Id	OneR	Id	ReliefF	Id	SymmetricalUncert	Id	ChiSquared
0	0,3121	24	0,0955	31	44,334	9	0,0226	24	0,0793	26	63,906
1	0,3059	23	0,0955	30	44,316	5	0,0183	23	0,0793	23	63,906
4	0,304	30	0,0728	6	41,945	30	0,0108	30	0,0572	31	57,157
2	0,2999	33	0,0727	40	41,945	33	0,0108	33	0,0571	30	57,106
11	0,2956	31	0,0727	38	41,945	31	0,0107	31	0,0571	33	57,022
23	0,2598	9	0,0606	37	41,945	6	0,0085	9	0,0442	9	35,162
24	0,2598	6	0,0435	10	41,631	21	0,0069	6	0,0376	6	34,508
26	0,2598	38	0,0321	21	40,777	37	0,0067	38	0,0279	21	24,080
19	0,2558	40	0,0321	3	40,603	40	0,0067	40	0,0279	40	22,934
17	0,1621	37	0,0321	1	40,533	38	0,0067	37	0,0279	37	22,934
6	0,1586	10	0,0246	4	40,516	20	0,0055	10	0,0220	38	22,934
30	0,1437	21	0,0197	2	40,481	10	0,0052	21	0,0188	10	20,445
33	0,1436	7	0,0165	19	40,463	3	0,0040	7	0,0150	3	18,028
31	0,1435	8	0,0150	11	40,446	7	0,0035	3	0,0141	1	15,613
25	0,1319	3	0,0147	7	40,411	2	0,0034	8	0,0135	7	15,405
27	0,1319	1	0,0127	8	40,394	13	0,0034	1	0,0122	4	15,010
10	0,1246	4	0,0122	17	40,271	17	0,0033	4	0,0118	8	14,029
38	0,122	2	0,0113	34	40,237	19	0,0029	2	0,0108	2	13,804
40	0,122	11	0,0103	32	40,237	4	0,0028	11	0,0099	11	12,600
37	0,122	19	0,0099	36	40,080	1	0,0026	19	0,0096	19	11,835
18	0,1202	17	0,0091	39	40,080	11	0,0023	17	0,0086	17	68,299
7	0,1005	20	0,0080	41	40,080	15	0,0022	20	0,0073	20	43,105
9	0,0864	18	0,0040	12	40,080	18	0,0021	18	0,0038	13	25,566
8	0,0822	13	0,0039	9	40,080	14	0,0020	13	0,0037	18	24,830
13	0,0775	15	0,0035	14	40,080	8	0,0017	15	0,0033	25	22,483
15	0,0766	27	0,0034	13	40,080	12	0,0008	27	0,0033	27	22,483
16	0,0763	25	0,0034	16	40,080	27	0,0006	25	0,0033	16	21,984
14	0,0625	16	0,0033	15	40,080	25	0,0006	16	0,0032	15	21,071
20	0,0481	14	0,0030	28	40,080	16	0,0001	14	0,0029	14	18,359
41	0	41	0	27	40,080	32	0,0001	41	0	41	0
39	0	39	0	29	40,080	34	0,0001	39	0	39	0
12	0	28	0	18	40,080	39	2,1169	28	0	32	0
35	0	36	0	35	40,080	41	2,1169	36	0	12	0
36	0	34	0	26	40,080	43	0	34	0	28	0
42	0	29	0	25	40,080	42	0	29	0	35	0
34	0	32	0	24	40,080	35	0	32	0	29	0
32	0	22	0	23	40,080	36	0	22	0	36	0
29	0	35	0	20	40,080	29	0	35	0	34	0
28	0	42	0	42	40,080	0	0	42	0	22	0
22	0	12	0	22	40,080	28	0	12	0	42	0
43	0	43	0	43	40,080	22	0	43	0	43	0

Posteriormente, se evaluó el modelo, usando las técnicas de clasificación con mejores resultados obtenidos del escenario tres (3), hibridando las siguientes técnicas de selección de características: *Gain Ratio, Info Gain OneR, Symmetrical Uncert* y *ChiSquared*.Ver tabla No. 16.

Tabla No. 16
Resultados Hibridación Técnicas de Clasificación y selección

Clasificador	# Atr	FPR	Precision	Recall	ROC area	Selección
Classification Via Regression	5	0.25%	99.95%	99.95%	99.96%	Gain Ratio
	5	0.25%	99.95%	99.95%	99.96%	Info Gain
	5	0.25%	99.95%	99.95%	99.96%	OneR
	5	0.25%	99.95%	99.95%	99.96%	Symmetrica lUncert
	5	0.25%	99.95%	99.95%	99.96%	ChiSquared
OneR	5	0.14%	99.75%	99.75%	99.80%	Gain Ratio
	5	0.14%	99.75%	99.75%	99.80%	Info Gain
	5	0.14%	99.75%	99.75%	99.80%	OneR
	5	0.14%	99.75%	99.75%	99.80%	Symmetrical Uncert
	5	0.14%	99.75%	99.75%	99.80%	ChiSquared

Fuente propia del autor

Para determinar cuál de estas técnicas de clasificación generó mejores resultados, se evaluaron las diferentes métricas de calidad.

Las técnicas de selección usadas no inciden ostensiblemente en generar un valor diferencial en las métricas ROC Area, *Recall* y *Precision*, su incidencia se refleja en determinar el número de características más apropiadas. Se evidencia una disminución del número de características y cuales tienen mayor incidencia en el modelo, pero no se obtuvo un resultado significativo aplicando las diferentes técnicas de selección. Sin embargo, al siminuir el numero de características mejoramos el tiempo computacional para la realización del modelo.
Se ratifica que los clasificadores *Classification Via Regression* y *OneR* con cinco (5) características con cualquiera de las técnicas de selección arrojan importantes resultados para

Classification Via Regression el ROC área es 99,96% y *Recall* 99.95%, para *OneR* el ROC área es 99,80% y *Recall* 99.75%. A continuación, se muestra un comparativo entre el clasificador *Classification Via Regression* y *OneR* para cada una de las clases. Ver Tabla 17.

Tabla No. 17

Resultado comparativo entre el clasificador Classification Via Regression + Gain ratio y OneR + Gain Ratio

clase	Classification Via Regression + Gain ratio (5 caracteristicas)				OneR + Gain ratio (5 caracteristicas)			
	FP-Rate	Precision	Recall	Roc area	FP-Rate	Precision	Recall	Roc area
Comer	0,00%	100%	100%	100%	0,0%	100%	99,80%	99.80%
Limpiar	0,00%	100%	100%	100%	0,1%	99,90%	99,90%	100%
Cocinar	1,00%	99,90%	99,90%	100%	0.3%	99,50%	99,70%	99,80%
Llamada Telefonica	0,00%	100%	100%	100%	0,0%	100%	100%	100%
Lavarse las manos	0,00%	100%	99,60%	99.40%	0,0%	100%	97.80%	97,90%
Total	0,25%	99.95%	99.95%	99,96%	0,14%	99.75%	99.75%	99,80%

Fuente propia del autor

6. CONCLUSIONES

Después de analizar cada uno de los escenarios se puede concluir que la segmentación y el balanceo de carga mejoraron los valores de las metricas de calidad a través de cada uno de los escenarios como se detalla en la tabla No. 18

Tabla No. 18

Comparativo de las métricas de calidad del escenario1, escenario 2 y escenario 3

Clasificador	Escenario 1 *Dataset* preprocesado		Escenario 2 *Dataset* segmentado		Escenario 3 *Dataset* segmentado y balanceado	
	FPR	Recall	FPR	Recall	FPR	Recall
Classification Via Regression	**1,20%**	**97,60%**	**0,10%**	**99,80%**	**0.28%**	**99.95%**
Random SubSpace	**4,10%**	**89,90%**	**3,80%**	**91,10%**	**2.15%**	**94.75%**
Bagging	5,20%	86,40%	3,50%	90,00%	3.42%	90.28%
RandomForest	4,10%	87,40%	3,20%	90,40%	3.10%	91.09%
Attribute Selected	2,70%	91,80%	0,90%	97,10%	0.69%	97.92%
J48	2,70%	91,60%	0,90%	96,80%	0.77%	97.64%
OneR	**2,80%**	**94,80%**	**0,30%**	**99,50%**	**0.14%**	**99.75%**
LMT	5,50%	85,20%	3,90%	88,60%	3,35%	89,67%
REP Tree	5,70%	84,50%	3,60%	89,20%	3,52%	89.54%
RandomCommittee	4,20%	87,20%	3,30%	89,90%	3,18%	90,64%

Fuente propia del autor

Como se evidencia para el primer escenario el clasificador *Classification Via Regression* tuvo un *Recall* de 97,60%, después de realizar el proceso de segmentación esta misma métrica tuvo un resultado de 99,80% y para el tercer escenario después del balanceo de cargas este mismo clasificador tuvo un *recall* de 99,95%. De igual manera podemos evidenciar una mejora en la métrica de FPR, para el primer escenario se obtuvo un resultado de 1,20% finalmente para el tercer escenario fue de 0,28%, con lo que podemos intuir que el proceso de segmentación y

balanceo de carga mejoran sustancialmente el desempreño del clasificador *Classification Via Regression.*

Cabe anotar que, el clasificador *OneR ob*tuvo mejores rendimientos a través de cada uno de los escenarios. Para el primero este clasificador se posicionaba en el séptimo lugar con un *Recall* de 94,80% y una tasa de FPR de 2,80%, lugo del proceso de segmentación se ubicó en el segundo lugar con un *Recall* de 99,50% y una tasa de FPR de 0,30%, finalmente luego del proceso de balanceo de cargas los resultados obtenidos fueron de 99,75% y 0,14% para cada una de las métricas.

En el cuarto escenario se realizó la priorización de los atributos y luego de la hibridación de las técnicas de clasificación y selección se estableció que el proceso más significativo con respecto a los mejores resultados obtenidos fue el de balanceo de carga, debido a que los atributos de mayor incidencia sobre el modelo son las que no poseen funciones de agregación las cuales son generadas por el proceso de segmentación a continuación, en la tabla 19 se muestra la priorización de los atributos (10 mejores) con la técnica de selección *Gain Ratio*, de las cuales las cinco (5) primeras son las usadas para la generación del modelo.

Tabla No. 19

Priorización de características según técnica de selección de atributos

Id	GainRatio	NombreAtr
21	**0,3294**	**asterisk**
5	**0,3173**	**M13**
3	**0,3133**	**M08**
0	**0,3121**	**DATE**
1	**0,3059**	**M01**
4	0,304	M09
2	0,2999	M07
11	0,2956	M23
23	0,2598	AD1-A-MAX
24	0,2598	AD1-A-MIN

Después de analizar los escenarios anteriores, también se realizó el estudio de las métricas de calidad del mejor clasificador con respecto a las clases para cada uno de los escenarios. Ver tabla No. 20

Tabla No. 20

Comparación de clase basado en Recall por cada uno de los escenarios

	Escenario 1		**Escenario 2**		**Escenario 3**	
Clase	**FPR**	**Recall**	**FPR**	**Recall**	**FPR**	**Recall**
Comer	0,10%	96,30%	0%	99,70%	0%	100%
Limpiar	0,10%	99,70%	0%	99,90%	0%	100%
Cocinar	2,70%	99,70%	0,30%	100%	0,10%	100%
Llamada Telefonica	0,50%	97,90%	0%	99,90%	0%	100%
Lavarse las manos	0,20%	71,20%	0%	99,80%	0%	99,20%

El balanceo de cargas favoreció considerablemente a la actividad lavarse las manos en la tabla 20 podemos observar que en un primer escenario donde no se había realizado balanceo de cargas el *Recall* es de 71,20%, en el escenario dos (2) luego del balanceo de cargas el *Recall* mejoró considerablemente a 99,80%, con estos resultados podemos decir que lavarse las manos fue la actividad que mejores beneficios tuvo después del balanceo de cargas.

7. Referencias

Amiribesheli, M., Benmansour, A., & Bouchachia, A. (2015). A review of smart homes in healthcare. *Journal of Ambient Intelligence and Humanized Computing*, *6*(4), 495–517. https://doi.org/10.1007/s12652-015-0270-2

Andrew McCallum, K. N. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization*, *752*, 307. https://doi.org/10.3115/1067807.1067848

Asaju, L. B., Shola, P. B., Franklin, N., & Abiola, H. M. (2017). Intrusion Detection System on a Computer Network Using an Ensemble of Randomizable Filtered Classifier, K-Nearest …. *Ftstjournal.Com*, *2*(1), 550–553. www.ftstjournal.com

Azevedo, M., & Santos, M. (2005). *Data Mining Descoberta de Conhecimento em bBase de Dados*.

Bidgoli, A., & Naseriparsa, M. (2018). A Hybrid Feature Selection by Resampling, Chi-squared abd Consistency Evaluation Techniques Flexible Interactive Querying View project. In *researchgate.net*. https://www.researchgate.net/publication/327106211

Björk, K.-M., Eirola, E., Miche, Y., & Lendasse, A. (2016). *A new application of machine learning in health care*. 1–4. https://doi.org/10.1145/2910674.2935861

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*. https://doi.org/10.1201/9781315139470

Burgueño, M. J., García-Bastos, J. L., & González-Buitrago, J. M. (1995). ROC curves in the evaluation of diagnostic tests. *Medicina Clínica*, *104*(17), 661–670.

Cai, Y. D., Feng, K. Y., Lu, W. C., & Chou, K. C. (2006). Using LogitBoost classifier to predict protein structural classes. *Journal of Theoretical Biology*, *238*(1), 172–176. https://doi.org/10.1016/j.jtbi.2005.05.034

Capela, N. A., Lemaire, E. D., & Baddour, N. (2015). Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients. *PLoS ONE*, *10*(4), e0124414. https://doi.org/10.1371/journal.pone.0124414

Cardoso, H. L., & Moreira, J. M. (2016). Human Activity Recognition by Means of Online Semi-supervised Learning. *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, 75–77. https://doi.org/10.1109/MDM.2016.93

Carlos, A., D'negri, E., De Vito, E. L., & Zadeh, L. A. (2006). Introducción al razonamiento aproximado: lógica difusa. In *Revista Argentina de Medicina Respiratoria Año* (Vol. 6).

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *ACM International Conference Proceeding Series*, *148*, 161–168. https://doi.org/10.1145/1143844.1143865

Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009). Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*. https://doi.org/10.1109/CBMS.2009.5255394

Chandra, S., & Maheshkar, S. (2017). Verification of static signature pattern based on random subspace, REP tree and bagging. *Multimedia Tools and Applications*, *76*(18), 19139–19171. https://doi.org/10.1007/s11042-017-4531-2

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

https://sci-hub.si/http://www.jair.org/index.php/jair/article/view/10302

Chen, L., Nugent, C. D., & Wang, H. (2012). A knowledge-driven approach to activity recognition in smart homes. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 961–974. https://doi.org/10.1109/TKDE.2011.51

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, *151*, 147–160. https://doi.org/10.1016/j.catena.2016.11.032

Chen, Y., & Shen, C. (2017). Performance Analysis of Smartphone-Sensor Behavior for Human Activity Recognition. *IEEE Access*, *5*, 3095–3110. https://doi.org/10.1109/ACCESS.2017.2676168

Choudhury, S., & Bhowal, A. (2015). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings*, 89–95. https://doi.org/10.1109/ICSTM.2015.7225395

Cook, D. J. (2006). Health monitoring and assistance to support aging in place. *Journal of Universal Computer Science*, *12*(1), 15–29. https://doi.org/10.3217/jucs-012-01-0015

Cook, D. J., Crandall, A., Singla, G., & Thomas, B. (2010). Detection of social interaction in smart spaces. *Cybernetics and Systems*, *41*(2), 90–104. https://doi.org/10.1080/01969720903584183

Cortes, C., Vapnik, V., & Saitta, L. (1995). Support-Vector Networks Editor. In *Machine Leaming* (Vol. 20). Kluwer Academic Publishers.

Crandall, A. S. (2011). *BEHAVIOMETRICS FOR MULTIPLE RESIDENTS IN A SMART ENVIRONMENT*. https://sci-hub.si/http://research.wsulibs.wsu.edu/xmlui/handle/2376/2855

Crandall, A. S., & Cook, D. J. (2013). *Behaviometrics for Identifying Smart Home Residents* (pp. 55–71). https://doi.org/10.2991/978-94-6239-018-8_4

Crandall, A. S., & Cook, D. J. (2010). Using a Hidden Markov Model for resident identification. *Proceedings - 2010 6th International Conference on Intelligent Environments, IE 2010*, 74–79. https://doi.org/10.1109/IE.2010.21

Daelemans, W., Hoste, V., De Meulder, F., & Naudts, B. (2003). Combined optimization of feature selection and algorithm parameters in machine learning of language. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, *2837*, 84–95. https://doi.org/10.1007/978-3-540-39857-8_10

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340–341*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033

Dietterich, T. G. (1997). Machine-Learning Research. In *aaai.org*. https://sci-hub.st/https://www.aaai.org/ojs/index.php/aimagazine/article/view/1324

Ding, D., Cooper, R. A., Pasquina, P. F., & Fici-Pasquina, L. (2011). Sensor technology for smart homes. *Maturitas*, *69*(2), 131–136. https://doi.org/10.1016/j.maturitas.2011.03.016

Du, W. S., & Hu, B. Q. (2014). Approximate distribution reducts in inconsistent interval-valued ordered decision tables. *Information Sciences*, *271*, 93–114. https://doi.org/10.1016/j.ins.2014.02.070

Eddy, S. R. (1998). Profile hidden Markov models. *Academic.Oup.Com*, *144*(9), 755–763. https://academic.oup.com/bioinformatics/article-abstract/14/9/755/259550

Envejecimiento y salud. (2018, February 5). https://www.who.int/es/news-room/fact-sheets/detail/envejecimiento-y-salud

Fahad, L. G., Tahir, S. F., & Rajarajan, M. (2015). Feature selection and data balancing for activity recognition in smart homes. *IEEE International Conference on Communications*, *2015-Septe*, 512–517. https://doi.org/10.1109/ICC.2015.7248373

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–53. https://vvvvw.aaai.org/ojs/index.php/aimagazine/article/view/1230

Fleury, A., Vacher, M., & Noury, N. (2010). SVM-Based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 274–283. https://doi.org/10.1109/TITB.2009.2037317

Gaikwad, N. B., Tiwari, V., Keskar, A., & Shivaprakash, N. C. (2019). Efficient FPGA Implementation of Multilayer Perceptron for Real-Time Human Activity Classification. *IEEE Access*, *7*, 26696–26706. https://doi.org/10.1109/ACCESS.2019.2900084

Galván-Tejada, C. E., Galván-Tejada, J. I., Celaya-Padilla, J. M., Delgado-Contreras, J. R., Magallanes-Quintanar, R., Martinez-Fierro, M. L., Garza-Veloz, I., López-Hernández, Y., & Gamboa-Rosales, H. (2016). An Analysis of Audio Features to Develop a Human Activity Recognition Model Using Genetic Algorithms, Random Forests, and Neural Networks. *Mobile Information Systems*, *2016*, 1–10. https://doi.org/10.1155/2016/1784101

Gudivada, V. N., Ding, J., & Apon, A. (2017). *Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations Flow Cytometry of 3-D structure View project Data Quality Considerations for Big Data and Machine*

Learning: Going Beyond Data Cleaning and Transf. *October*, 1–20. https://www.researchgate.net/publication/318432363

Han, S., Cao, Q., & Han, M. (2012). Parameter selection in SVM with RBF kernel function. *World Automation Congress Proceedings*. https://ezproxy.cuc.edu.co:2065/document/6321759

He, Y., Li, Y., & Yin, C. (2012). Falling-Incident Detection and Alarm by Smartphone with Multimedia Messaging Service (MMS). *E-Health Telecommunication Systems and Networks*, *01*(01), 1–5. https://doi.org/10.4236/etsn.2012.11001

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Hoey, J., Pltz, T., Jackson, D., Monk, A., Pham, C., & Olivier, P. (2011). Rapid specification and automated generation of prompting systems to assist people with dementia. *Pervasive and Mobile Computing*, *7*(3), 299–318. https://doi.org/10.1016/j.pmcj.2010.11.007

Islam, A. (2018). Android Application Based Smart Home Automation System Using Internet of Things. *2018 3rd International Conference for Convergence in Technology (I2CT)*, 1–9. https://doi.org/10.1109/I2CT.2018.8529752

Jalal, A., Kamal, S., & Kim, D. (2017). A Depth Video-based Human Detection and Activity Recognition using Multi-features and Embedded Hidden Markov Models for Health Care Monitoring Systems. *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(4), 54. https://doi.org/10.9781/ijimai.2017.447

Jiang, X., Liu, M., Zheng, D., Zhao, X., Wang, W., & Sun, G. (2017). Enhanced photocatalytic degradation of organic pollutants using high-aspect-ratio Si/ITO/WO3 micropost photoelectrodes. *2017 IEEE 30th International Conference on Micro Electro Mechanical*

Systems (MEMS), 881–884. https://doi.org/10.1109/MEMSYS.2017.7863549

Kalmegh, S. (2015). Analysis of WEKA Data Mining Algorithm REPTree , Simple Cart and RandomTree for Classification of Indian News. *International Journal of Innovative Science, Engineering & Technology*, *2*(2), 438–446. www.ijiset.com

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271–277. https://sci-hub.st/http://csjournals.com/IJITKM/PDF 3-1/19.pdf

Khalajzadeh, H., Mansouri, M., & Teshnehlab, M. (2014). Face recognition using convolutional neural network and simple logistic classifier. *Advances in Intelligent Systems and Computing*, *223*, 197–207. https://doi.org/10.1007/978-3-319-00930-8_18

Khalifa, S., Lan, G., Hassan, M., Seneviratne, A., & Das, S. K. (2018). HARKE: Human Activity Recognition from Kinetic Energy Harvesting Data in Wearable Devices. *IEEE Transactions on Mobile Computing*, *17*(6), 1353–1368. https://doi.org/10.1109/TMC.2017.2761744

Kira, K., & Rendell, L. A. (1992). Feature selection problem: traditional methods and a new algorithm. *Proceedings Tenth National Conference on Artificial Intelligence*, 129–134. https://sci-hub.st/https://www.aaai.org/Library/AAAI/1992/aaai92-020.php

Kumar, K., Kumar, G., & Kumar, Y. (2013). Feature Selection Approach for Intrusion Detection System. *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, *2*(5), 47–53. http://warse.org/pdfs/2013/icceitsp09.pdf

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *59*(1–2), 161–205. https://doi.org/10.1007/s10994-005-0466-3

Lladó, M. R., Código, H., Lennin, A., Quiroz, P., & Lima -Perú, V. (2020). ENTORNO

DOMÓTICO ADAPTADO A PERSONAS CON DISCAPACIDAD FÍSICA UTILIZANDO MODELOS OCULTOS DE MARKOV Tesis para optar el Título Profesional de Ingeniero de Sistemas. In *Repositorio Institucional - Ulima*. Universidad de Lima. http://repositorio.ulima.edu.pe/handle/20.500.12724/11664

López de Ullibarri, G. I., & Píta Fernández, S. (1998). Curvas ROC. *Cad Aten Primaria*, *5*(4), 229–235.

López Saca, F., Ferreyra Ramírez, A., Avilés Cruz, C., Villegas Cortez, J., Zúñiga López, A., & Rodrigez Martinez, E. (2018). Preprocesamiento de bases de datos de imágenes para mejorar el rendimiento de redes neuronales convolucionales. *Research in Computing Science*, *147*(7), 35–45. https://doi.org/10.13053/rcs-147-7-3

Marcondes, C. H., & Almeida Campos, M. L. de. (2008). ONTOLOGIA E WEB SEMÂNTICA: O ESPAÇO DA PESQUISA EM CIÊNCIA DA INFORMAÇÃO. *PontodeAcesso*, *2*(1), 107. https://doi.org/10.9771/1981-6766rpa.v2i1.2669

Mejia-Ricart, L. F., Helling, P., & Olmsted, A. (2018). Evaluate action primitives for human activity recognition using unsupervised learning approach. *2017 12th International Conference for Internet Technology and Secured Transactions, ICITST 2017*, 186–188. https://doi.org/10.23919/ICITST.2017.8356374

Mishra, A. (2018). *Metrics to Evaluate your Machine Learning Algorithm*.

Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, *5*(6), 865–872. https://doi.org/10.1109/72.329683

Nagalla, R., Pothuganti, P., & Pawar, D. S. (2017). Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random

Forests. *Procedia Computer Science*, *109*, 474–481. https://doi.org/10.1016/j.procs.2017.05.312

Nagi, S. Z., Burk, R. D., & Potter, H. R. (1965). Back disorders and rehabilitation achievement. *Journal of Chronic Diseases*, *18*(2), 181–197. https://doi.org/10.1016/0021-9681(65)90101-3

Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, *33*(4), 275–306. https://doi.org/10.1007/s10462-010-9156-z

Nuraeni, F., Febriani Sm, N. N., Listiani, L., & Rahmawati, E. (2019). Implementation of K-Means Algorithm with Distance of Euclidean Proximity in Clustering Cases of Violence Against Women and Children. *2019 1st International Conference on Cybernetics and Intelligent System, ICORIS 2019*, 162–167. https://doi.org/10.1109/ICORIS.2019.8874883

Parimala, R., & Nallaswamy, R. (2011). A Study of Spam E-mail classification using Feature Selection package. *Global Journal of Computer Science and Technology*, *11*(7), 45–54. https://computerresearch.org/index.php/computer/article/view/727

Qian, H., Mao, Y., Xiang, W., & Wang, Z. (2010). Recognition of human activities using SVM multi-class classifier. *Pattern Recognition Letters*, *31*(2), 100–111. https://doi.org/10.1016/j.patrec.2009.09.019

Rajput, A., Aharwal, R. P., Dubey, M., Saxena, S. P., & Raghuvanshi, M. (2011). J48 and JRIP rules for e-governance data. *International Journal of Computer Science and Security*, *5*(2), 201–207.

Ren, X., & Malik, J. (2003). Learning a classification model for segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, *1*, 10–17.

https://doi.org/10.1109/iccv.2003.1238308

Ricardo, S., Pereira Isabel Hernández Arteaga Segundo, T., Caicedo Zambrano Arsenio, J., Troya, H., & Carlos Alvarado Pérez, J. (2015). *DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO ACADÉMICO*. https://sci-hub.st/https://repository.ucc.edu.co/handle/20.500.12494/1039

Ronao, C. A., & Cho, S. B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, *59*, 235–244. https://doi.org/10.1016/j.eswa.2016.04.032

Ruan, Y. X., Lin, H. T., & Tsai, M. F. (2014). Improving ranking performance with cost-sensitive ordinal classification via regression. *Information Retrieval*, *17*(1), 1–20. https://doi.org/10.1007/s10791-013-9219-2

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, *16*(3), 235–240. https://doi.org/10.1007/bf00993309

Seelye, A., Schmitter-Edgecombe, M., Cook, D. J., & Crandall, A. (2013). Smart environment prompting technologies for everyday activities in mild cognitive impairment. *Journal of the International Neuropsychological Society*, *19*, 442–452.

Shah, C. (2020). Supervised Learning. In *A Hands-On Introduction to Data Science* (pp. 235–289). https://doi.org/10.1017/9781108560412.010

Spruit, M., Vroon, R., & Batenburg, R. (2014). Towards healthcare business intelligence in long-term care: An explorative case study in the Netherlands. *Computers in Human Behavior*, *30*, 698–707. https://doi.org/10.1016/j.chb.2013.07.038

Suykens, J. A. K., & Vandewalle, J. (1999). Training multilayer perceptron classifiers based on a

modified support vector method. *IEEE Transactions on Neural Networks*, *10*(4), 907–911. https://doi.org/10.1109/72.774254

Tabuenca Dopico, P., Sánchez Espeso, P. P., & Villar Bonet, E. (1993). Realización de un planificador inteligente para síntesis de alto nivel. *VIII Congreso Diseño de Circuitos Integrados: Málaga, 9 Al 11 de Noviembre de 1993, 1993, Págs. 315-319*, 315–319. https://dialnet.unirioja.es/servlet/articulo?codigo=6418065&info=resumen&idioma=SPA

Universidad Pedagógica y Tecnológica de Colombia. (2004). *Unidad 1 Estadistica Descriptiva*.

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064

Yiyu, Y. (2007). Decision-theoretic rough set models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4481 LNAI*, 1–12. https://doi.org/10.1007/978-3-540-72458-2_1

Yuan, G., Wang, Z., Meng, F., Yan, Q., & Xia, S. (2019). An overview of human activity recognition based on smartphone. *Sensor Review*, *39*(2), 288–306. https://doi.org/10.1108/SR-11-2017-0245